The Effect of Metadata on Scientific Literature Tagging: A Cross-Field Cross-Model Study

Yu Zhang, Bowen Jin, Qi Zhu, Yu Meng, Jiawei Han
University of Illinois at Urbana-Champaign
{yuz9, bowenj4, qiz3, yumeng5, hanj}@illinois.edu

ABSTRACT

Due to the exponential growth of scientific publications on the Web, there is a pressing need to tag each paper with fine-grained topics so that researchers can track their interested fields of study rather than drowning in the whole literature. Scientific literature tagging is beyond a pure multi-label text classification task because papers on the Web are prevalently accompanied by metadata information such as venues, authors, and references, which may serve as additional signals to infer relevant tags. Although there have been studies making use of metadata in academic paper classification, their focus is often restricted to one or two scientific fields (e.g., computer science and biomedicine) and to one specific model. In this work, we systematically study the effect of metadata on scientific literature tagging across 19 fields. We select three representative multi-label classifiers (i.e., a bag-of-words model, a sequence-based model, and a pre-trained language model) and explore their performance change in scientific literature tagging when metadata are fed to the classifiers as additional features. We observe some ubiquitous patterns of metadata's effects across all fields (e.g., venues are consistently beneficial to paper tagging in almost all cases), as well as some unique patterns in fields other than computer science and biomedicine, which are not explored in previous studies.

CCS CONCEPTS

• **Information systems** → **Digital libraries and archives**; **Data mining**; **World Wide Web**.

KEYWORDS

scientific literature tagging; metadata; text classification

ACM Reference Format:
Yu Zhang, Bowen Jin, Qi Zhu, Yu Meng, Jiawei Han. 2023. The Effect of Metadata on Scientific Literature Tagging: A Cross-Field Cross-Model Study. In *Proceedings of the ACM Web Conference 2023 (WWW '23), May 1–5, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3543507.3583354

1 INTRODUCTION

A variety of academic service platforms, such as Google Scholar, AMiner [40], Microsoft Academic [38], Semantic Scholar [1], and PubMed [26], are available on the Web with great attention received. One major goal of these platforms is to help researchers query and track academic information and resources. Meanwhile, the volume of scientific publications is growing exponentially, doubling every 12 years [12] and reaching 240,000,000 by 2019 [42]. In such an information explosion era, it becomes more important than ever to accurately tag each scientific paper with its relevant topics so that researchers can track their interested fields of study instead of getting overwhelmed by the whole literature. Figure 1 shows an example of the scientific literature tagging task, which aims to predict the tags such as "World Wide Web", "Webgraph", and "Link Farm" given the paper "*Graph structure in the Web*".

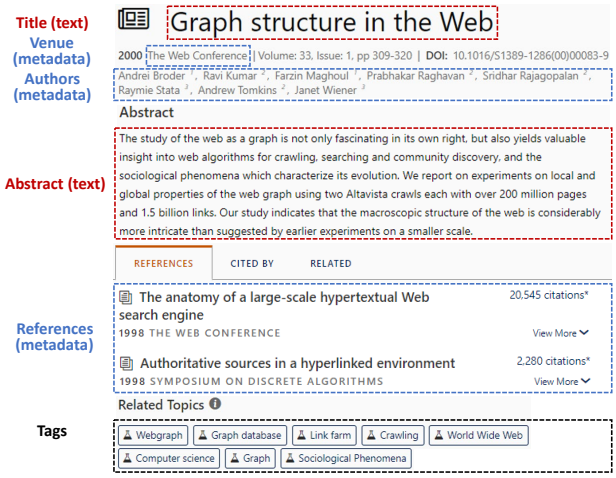


Figure 1: A scientific paper with metadata from Microsoft Academic [38]. The goal of scientific literature tagging is to predict its related topics.

Previous studies [47, 59] have pointed out that scientific literature tagging is beyond a multi-label text classification task because academic papers are accompanied by metadata, which make them more complex than plain text sequences. Such metadata information, including venues, authors, and references, can be strong indicators of each paper's related topics. For example, in Figure 1, the venue "*WWW*" implies the paper's relevance to "Computer Science" and "World Wide Web", while the authors and references may further indicate fine-grained tags such as "Webgraph" and "Link Farm".

Although existing studies on scientific literature tagging [45, 47, 57, 59, 60] have proposed to incorporate metadata features into the tagger, they still have two major limitations. First, their focus is restricted to one or two scientific fields only (e.g., computer science and biomedicine). Empirically examining the effect of metadata in other fields (e.g., art, economics, mathematics, physics, etc.) has remained elusive, mainly owing to the lack of benchmark datasets for fine-grained paper tagging in these fields. Second, the proposed

[†]Code and Datasets are available at https://github.com/yuzhimanhua/MAPLE and https://doi.org/10.5281/zenodo.7611544, respectively.

WWW '23, May 1–5, 2023, Austin, TX, USA

ACM ISBN 978-1-4503-9416-1/23/04…$15.00
https://doi.org/10.1145/3543507.3583354

metadata-aware taggers mainly train an RNN [5] or Transformer [41] architecture from scratch. It is still unclear whether the proposed usage of metadata can be generalized to other approaches, such as bag-of-words classifiers (which are still broadly studied in large-scale multi-label text classification when efficiency is concerned [23]) and pre-trained language models.

Contributions. To address the aforementioned two limitations of previous studies, in this work, we conduct a systematic *cross-field cross-model* study on the effect of metadata on scientific literature tagging. First, we construct a large-scale scientific literature tagging benchmark, MAPLE (Metadata-Aware Paper colLEction), from the Microsoft Academic Graph [38]. MAPLE covers 19 scientific fields and consists of more than 11.9 million papers. The number of candidate tags in each field ranges between ~700 and ~64,000. Then, we consider three major types of multi-label classification approaches: bag-of-words classifiers [2, 16, 19, 31–34, 39, 49, 50, 54], sequence-based classifiers [21, 44, 45, 47, 53, 59], and pre-trained language models [4, 17, 48, 52, 56, 60]. We select one representative model from each of the three categories, namely Parabel [33], Transformer [44], and OAG-BERT [24], that can be modified in a straightforward way to jointly take text and metadata as input for classification. Based on MAPLE, we explore the effect of venues, authors, and references on paper tagging in the 19 fields when using the three selected classifiers. We have the following major observations:

- The effect of metadata varies significantly across different fields and classifiers. In general, venues are consistently beneficial in almost all cases, while the benefit of authors and references is highly dependent on the field and the classifier. For example, author information is evidently beneficial to paper tagging in the Philosophy field while harmful in Geology; references are helpful in Mathematics when we use Parabel but become disadvantageous in the same field when Transformer is adopted.
- The effect of metadata tends to be similar in two fields that belong to the same high-level scientific area. For example, Biology and Medicine are both life sciences (according to [35, 51]), and the effects of venues, authors, and references are largely aligned in the two fields. This finding implies that the experience of using metadata in one field can be extrapolated to a similar field.
- We also study the effect of metadata when predicting tags at different granularity levels. In a number of fields, venues improve the performance for not only coarse-grained tags but also very fine-grained ones, which may be unusual in the Computer Science field. The major reason is that some venues in these fields can indicate very fine tags (e.g., "*Journal of Roman Archaeology*" in History, "*Mediaeval Studies*" in Philosophy).

To summarize, this work makes the following contributions: (1) We construct a large-scale benchmark, MAPLE, for scientific literature tagging across 19 fields, whose field coverage is much broader than the datasets used in previous paper tagging studies [45, 47, 59]. (2) We comprehensively evaluate the performance of different types of multi-label classifiers in scientific literature tagging after incorporating metadata features. (3) Our empirical findings demonstrate some ubiquitous patterns of metadata's effects across all fields, as well as some unique patterns in fields other than computer science and biomedicine, which are not explored in previous studies. Our systematic studies are meant for providing insights to practitioners to build scientific literature taggers that can benefit all fields.

Table 1: Statistics of the 20 datasets in MAPLE across 19 fields. There are 2 datasets in the Computer Science field, one of which is collected from top conferences and the other from top journals.

Field	Paper Source	#Papers	#Labels	#Venues	#Authors	#References
Art	Journal	58,373	1,990	98	54,802	115,343
Philosophy	Journal	59,296	3,758	98	36,619	198,010
Geography	Journal	73,883	3,285	98	157,423	884,632
Business	Journal	84,858	2,392	97	100,525	685,034
Sociology	Journal	90,208	1,935	98	85,793	842,561
History	Journal	113,147	2,689	99	84,529	284,739
Political Science	Journal	115,291	4,990	98	93,393	480,136
Environmental Science	Journal	123,945	694	100	265,728	1,217,268
Economics	Journal	178,670	5,205	97	135,247	1,042,253
Engineering	Journal	270,006	10,683	100	430,046	1,867,276
Psychology	Journal	372,954	7,641	100	460,123	2,313,701
Computer Science	Conference	263,393	13,613	75	331,582	1,084,440
	Journal	410,603	15,540	96	634,506	2,751,996
Geology	Journal	431,834	7,883	100	471,216	1,753,762
Mathematics	Journal	490,551	14,271	98	404,066	2,150,584
Materials Science	Journal	1,337,731	6,802	99	1,904,549	5,457,773
Physics	Journal	1,369,983	16,664	91	1,392,070	3,641,761
Biology	Journal	1,588,778	64,267	100	2,730,547	7,086,131
Chemistry	Journal	1,849,956	35,538	100	2,721,253	8,637,438
Medicine	Journal	2,646,105	36,619	100	4,345,385	7,405,779

2 PRELIMINARIES

Text and Metadata. We represent the text information of a paper p as a single text sequence $\mathcal{T}_p = w_1 w_2 ... w_{|\mathcal{T}_p|}$ by concatenating its title and abstract. The metadata of a paper p is represented as a set $\mathcal{M}_p = \{m_1, m_2, .., m_{|\mathcal{M}_p|}\}$ consisting of its venue, author(s), and reference(s).

Problem Definition. The scientific literature tagging task can be cast as a large-scale multi-label classification problem, where all candidate tags (e.g., "Organic Chemistry", "Coupling Reaction", "Suzuki Reaction") constitute a large label space $\mathcal{L}$ (e.g., with 10^3–10^5 labels). Given a scientific paper with its text and metadata information, the task is to find a set of labels from $\mathcal{L}$ that are relevant to the paper. Formally, the problem is defined as follows.

Definition 2.1. (Problem Definition) Given a training corpus $\mathcal{D}$ and the label space $\mathcal{L}$, where each paper $p \in \mathcal{D}$ is associated with its text $\mathcal{T}_p$, metadata information $\mathcal{M}_p$, and relevant tags $\mathcal{L}_p \subseteq \mathcal{L}$, our objective is to learn a multi-label text classifier f_{class} that can map an unlabeled paper $p' \notin \mathcal{D}$ to its relevant tags $\mathcal{L}_{p'} \subseteq \mathcal{L}$.

3 DATASET CONSTRUCTION

Previous studies on scientific literature tagging [9, 31, 45, 47, 52, 59] mainly use computer science and biomedicine papers to evaluate their proposed model, meanwhile paying less attention to other scientific fields. To bridge this gap, we construct MAPLE, a multi-field benchmark for evaluating scientific literature tagging. MAPLE is built upon data from the Microsoft Academic Graph (MAG) [38], which has been widely adopted in scientific text mining [6, 51, 59]. MAG covers 19 academic fields, which are listed in Table 1. For each field, we conduct the following steps to construct a dataset.

Venue Selection. MAG maintains a list of the top-100 journals in each field according to the h-index [13]. When constructing MAPLE, we focus on papers published in these top journals. Note that some preprint services (e.g., "*arxiv*", "*bioRxiv*", "*SSRN*", "*NBER Working Paper*") [43] are also viewed as top journals in MAG, but we exclude

them from our consideration. As a result, in Table 1, some of the constructed datasets contain less than 100 venues. Among the 19 fields, computer science (CS) has a unique publication culture: CS papers often appear first or exclusively in conferences rather than journals. Thus, for the Computer Science field, besides a collection of journal papers, we construct another dataset with papers from 75 top conferences according to CSRankings[1]. That being said, we will construct 20 datasets in total for the 19 fields.

Label Space Construction. For each field, we need a set of candidate labels for paper tagging. MAG has a directed acyclic graph (DAG)-structured label taxonomy $\mathcal{L}_{\text{MAG}}$ [37]. The taxonomy has 6 levels and more than 10^5 labels, where each of the 19 fields is a Layer-0 label (i.e., the most coarse-grained label). Given a field F, we extract its descendant labels in the taxonomy $\mathcal{L}_{\text{MAG}}$ as the label space $\mathcal{L}_F$ of this field, but we exclude F itself from $\mathcal{L}_F$ because the root label is trivial to predict in classification. Since a label may have more than one parent in the DAG-structured taxonomy, it may appear in the label space of two or more fields. For example, the label "Anatomy" is a child of both "Biology" and "Medicine", so it is a candidate label in both the Biology and the Medicine datasets in MAPLE.

Paper Selection. Each paper p in MAG is tagged with its relevant labels $\mathcal{L}_p \subseteq \mathcal{L}_{\text{MAG}}$ [37][2]. To be included in MAPLE (given a field F), a paper p needs to satisfy the following two criteria: (1) p is published in a selected venue of F; (2) p is labeled with F and at least one of F's descendants (i.e., $F \in \mathcal{L}_p$ and $|\mathcal{L}_p \cap \mathcal{L}_F| \geq 1$). When studying the scientific literature tagging task in a field F, we focus on labels related to that field only. Therefore, the ground truth labels of p are defined as $\mathcal{L}_{p|F} = \mathcal{L}_p \cap \mathcal{L}_F$. Note that a paper may appear in more than one field, and its ground truth labels are different when we consider different fields. For example, if a paper is tagged with "Medicine", "Polypharmacy", "Computer Science", and "Graph Embedding", given that "Polypharmacy" is a candidate label in the Medicine field and "Graph Embedding" is a candidate label in the Computer Science field, the paper will appear in both the Medicine and the Computer Science datasets in MAPLE. However, when we perform tagging in the Medicine field, the ground-truth label of the paper is "Polypharmacy"; when we perform tagging in the Computer Science field, the ground-truth label of the paper is "Graph Embedding".

Text and Metadata Extraction. For each selected paper p, we extract its title, abstract, venue, author(s), and reference(s) from MAG. The title and abstract are concatenated as text information $\mathcal{T}_p$. The venue, author(s), and reference(s) constitute metadata features $\mathcal{M}_p$ of the paper.

Training-Validation-Testing Split. MAPLE contains academic papers published between Jan. 1, 1981 and Dec. 31, 2020. We use papers from 1981 to 2015 for training and validation, and papers from 2016 to 2020 for testing.

Statistics of the constructed 20 datasets in MAPLE can be found in Table 1. More details are shown in Appendix A.1. From now on, for convenience of discussion, we use the terms "field" and "dataset" interchangeably if there is no ambiguity. (In other words, we treat "Computer Science (Conference)" and "Computer Science (Journal)" as different fields.)

[1]https://csrankings.org/

[2]The paper tags $\mathcal{L}_p$ come from the predictions of a system [37] proposed by Microsoft Academic. The tags are accurate as checked by humans [37] and have been used to support notable findings [18, 51]. Meanwhile, we also conduct experiments on three datasets with MeSH labels [7], which are curated by biomedical experts. Discussions on the additional three datasets can be found in Appendix A.5.

4 MODELS

Large-scale multi-label text classification (LMTC) has been extensively studied over the past decade. Various approaches are proposed and can be applied to scientific literature tagging. Based on how text is used as features in the classifier, existing LMTC approaches can be divided into three major categories: (1) *Bag-of-words classifiers* [2, 16, 19, 31–34, 39, 49, 50, 54] treat each document as a multiset of tokens while disregarding word position and order. Trees, embeddings, and linear layers are commonly used in these models. (2) *Sequence-based classifiers* [21, 44, 45, 47, 53, 59] take each document as a sequence of tokens and train a CNN, RNN, or Transformer architecture from scratch to build a multi-label classifier. (3) *Pre-trained language model classifiers* [4, 17, 48, 52, 56, 60] aim at transferring the knowledge learned from web-scale corpora (e.g., Wikipedia and PubMed) to the LMTC task, which can complement the text information from the training corpus. Since the goal of this paper is to study the effect of metadata, we select one LMTC model from each of the three categories that can be easily augmented with metadata information. Now we introduce the three selected models – Parabel [33] (bag-of-words), Transformer [44] (sequence-based), and OAG-BERT [24] (pre-trained language model) – and how they can take text and metadata features as input for classification.

4.1 Bag-of-Words Classifier: Parabel [33]

4.1.1 Using Text Only. In general, bag-of-words classifiers represent each document p as a $|\mathcal{V}_\mathcal{D}|$-dimensional vector $\boldsymbol{x}_p$, where $\mathcal{V}_\mathcal{D}$ is the vocabulary of the training corpus $\mathcal{D}$. Given a word $w \in \mathcal{V}_\mathcal{D}$, its corresponding entry in $\boldsymbol{x}_p$ is defined using the tf–idf score:

$$x_{p,w} = \text{tf}(w, p) \cdot \text{idf}(w, \mathcal{D}). \tag{1}$$

Here, $\text{tf}(w, p)$ is the term frequency of w in document p; $\text{idf}(w, \mathcal{D}) = \log \frac{|\mathcal{D}|}{|\{p' \in \mathcal{D}:\, w \in \mathcal{T}_{p'}\}|}$ is the inverse document frequency of w.

The relevant labels of each document are represented by an $|\mathcal{L}|$-dimensional vector $\boldsymbol{y}_p$, where $y_{p,l} = 1$ if l is a tag relevant to p (i.e., $l \in \mathcal{L}_p$), and $y_{p,l} = 0$ otherwise.

To perform multi-label text classification, Parabel learns an ensemble of multiple label trees, each of which is obtained by recursively partitioning the labels into two balanced groups until each node contains less than a certain number of labels. The partition process is implemented by spherical 2-means clustering based on the label representation $\boldsymbol{v}_l$, which is a unit vector in the direction of the mean of the training points containing label l. After label space partitioning, Parabel learns a hierarchical discriminative classifier $\Pr(\boldsymbol{y}_p|\boldsymbol{x}_p)$. Specifically, at each non-leaf node, a distribution is learned to determine which child nodes should be traversed; at each leaf node, a distribution is learned to predict the set of relevant tags. For more technical details, one can refer to [33], but we omit them here as they are not directly related to the usage of metadata.

4.1.2 Using Text + Metadata. It is straightforward to generalize bag-of-words classifiers to take metadata features. Let $\mathcal{U}_\mathcal{D}$ denote the set of metadata instances appearing in $\mathcal{D}$. Given a metadata instance $m \in \mathcal{U}_\mathcal{D}$ and a paper p, we define $\text{tf}(m, p)$ and $\text{idf}(m, \mathcal{D})$ as follows:

$$\text{tf}(m, p) = \mathbf{1}(m \in \mathcal{M}_p), \quad \text{idf}(m, \mathcal{D}) = \log \frac{|\mathcal{D}|}{|\{p' \in \mathcal{D} : m \in \mathcal{M}_{p'}\}|}, \tag{2}$$

where $\mathbf{1}(\cdot)$ is the indicator function. One can find that the definitions in Eq. (2) are well aligned with the definitions of $\text{tf}(w, p)$ and $\text{idf}(w, \mathcal{D})$, except that a metadata instance does not appear multiple times in a document.

Now we can have a "bag-of-metadata" representation $\tilde{\boldsymbol{x}}_p$ for each paper. $\tilde{\boldsymbol{x}}_p$ is a $|\mathcal{U}_\mathcal{D}|$-dimensional vector, where

$$\tilde{x}_{p,m} = \text{tf}(m, p) \cdot \text{idf}(m, \mathcal{D}). \tag{3}$$

Finally, we represent each paper p as a $(|\mathcal{U}_\mathcal{D}| + |\mathcal{V}_\mathcal{D}|)$-dimensional vector, that is, the concatenation of bag-of-words and "bag-of-metadata" representations $\boldsymbol{x}_p \,||\, \tilde{\boldsymbol{x}}_p$. This vector is fed into Parabel for training and prediction.

4.2 Sequence-based Classifier: Transformer [44]

4.2.1 Using Text Only. To apply Transformer [41] to LMTC, we follow the architecture proposed in [44], which adds multiple "[CLS]" tokens in front of document text $\mathcal{T}_p$ as the input sequence. Formally, given $\mathcal{T}_p = w_1 w_2 ... w_{|\mathcal{T}_p|}$, the input sequence of paper p is

$$\mathcal{I}_p = [\text{CLS}_1]\ [\text{CLS}_2]\ ...\ [\text{CLS}_C]\ w_1\ w_2\ ...\ w_{|\mathcal{T}_p|} \tag{4}$$

The motivation here is that when the label space is large (e.g., with 10^4 tags), the output representation of one "[CLS]" token (e.g., a vector with several hundred dimensions) may not carry enough information to predict relevant labels. Therefore, multiple "[CLS]" tokens are needed to probe the semantics of text from different perspectives.

After Transformer encodes the input sequence $\mathcal{I}_p$, each token $w \in \mathcal{I}_p$ will have an output representation $\boldsymbol{h}_w$. We concatenate the representations of all "[CLS]" tokens together as the paper embedding:

$$\boldsymbol{h}_p = \boldsymbol{h}_{[\text{CLS}_1]} \,||\, \boldsymbol{h}_{[\text{CLS}_2]} \,||\, ... \,||\, \boldsymbol{h}_{[\text{CLS}_C]}. \tag{5}$$

$\boldsymbol{h}_p$ is further fed into a fully connected layer to perform multi-label classification:

$$\hat{\boldsymbol{y}}_p = \text{Sigmoid}(\boldsymbol{W}^\top \boldsymbol{h}_p + \boldsymbol{b}). \tag{6}$$

Here, $\hat{\boldsymbol{y}}_p$ is an $|\mathcal{L}|$-dimensional vector, in which $\hat{y}_{p,l}$ is the predicted probability that paper p is relevant to label l. The classifier is trained to minimize the following binary cross-entropy (BCE):

$$-\sum_{l \in \mathcal{L}} (y_{p,l} \log \hat{y}_{p,l} + (1 - y_{p,l}) \log(1 - \hat{y}_{p,l})). \tag{7}$$

4.2.2 Using Text + Metadata. The fully connected attention mechanism in Transformer paves an easy way to incorporate metadata. To be specific, following [59], we can directly insert metadata tokens into the input sequence. Given paper text $\mathcal{T}_p = w_1 ... w_{|\mathcal{T}_p|}$ and metadata $\mathcal{M}_p = \{m_1, ..., m_{|\mathcal{M}_p|}\}$, the input sequence is

$$\widetilde{\mathcal{I}}_p = [\text{CLS}_1]\ ...\ [\text{CLS}_C]\ m_1\ ...\ m_{|\mathcal{M}_p|}\ [\text{SEP}]\ w_1\ ...\ w_{|\mathcal{T}_p|} \tag{8}$$

Unlike in CNN [21] or RNN [45, 53] classifiers, in Transformer, the order of metadata instances does not matter much because each pair of (metadata, word) or (metadata, metadata) can interact with each other via the fully connected attention mechanism during encoding. In our experiments, when listing authors in $\widetilde{\mathcal{I}}_p$, we follow the authorship order (i.e., 1st author, 2nd author, ...); when listing references in $\widetilde{\mathcal{I}}_p$, we adopt a random order because the citation order and inline contexts are not stored in MAG.

Following [59], we treat each metadata instance $m_i \in \widetilde{\mathcal{I}}_p$ as one token during Transformer encoding. For example, the venue "*WWW*" is represented as one token "[VENUE_1135342153]" instead of its textual name "*the web conference*" containing multiple tokens.

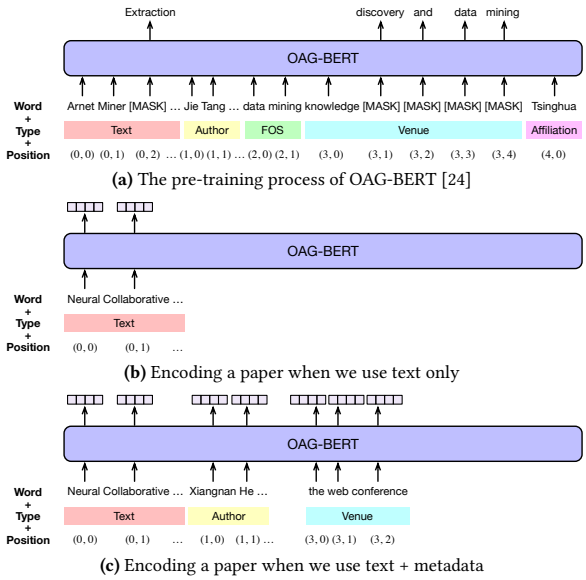


Figure 2: The pre-training process and our usage of OAG-BERT [24].

After $\widetilde{\mathcal{I}}_p$ is fed to Transformer, the remaining designs of the metadata-aware classifier exactly follow Eqs. (5)-(7).

4.3 Pre-trained Language Model Classifier: OAG-BERT [24]

Previous classifiers built upon pre-trained language models (PLMs) mainly utilize BERT [10], BioBERT [20], SciBERT [3], XLNet [46], or RoBERTa [25] to derive document representations. However, these PLMs mostly focus on text information during pre-training and are not specifically designed to deal with metadata. To bridge this gap, we adopt OAG-BERT [24], an entity-augmented academic language model, which can jointly encode scientific text and venue/author information.

The pre-training process of OAG-BERT is briefly illustrated in Figure 2(a). It places text and metadata information in a single sequence for masked language modeling. Besides, it proposes three strategies to deal with metadata entities: (1) heterogeneous entity type embedding makes the model aware of different metadata types; (2) span-aware entity masking selects a continuous span within long entities (e.g., the venue "*knowledge discovery and data mining*"); (3) 2-dimensional positional embedding jointly models inter and intra-entity token orders. The model is first trained on 5 million full-text papers and then on 120 million paper titles/abstracts with metadata from the Open Academic Graph [55].

4.3.1 Using Text Only. When considering paper text $\mathcal{T}_p$ only for classification, we first use OAG-BERT to encode the text sequence (as illustrated in Figure 2(b)).

$$\boldsymbol{H}_p = \text{OAG-BERT}(\mathcal{T}_p). \tag{9}$$

Here, $\boldsymbol{H}_p = [\boldsymbol{h}_{p,1}, ..., \boldsymbol{h}_{p,N}]$ contains the output vectors of all tokens. We then adopt mean pooling to obtain the paper representation $\boldsymbol{x}_p$.

$$\boldsymbol{x}_p = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{h}_{p,i}. \tag{10}$$

Directly fine-tuning the PLM using a BCE loss (i.e., Eq. (7)) is very time-consuming, especially when the label space is large. Considering efficiency and model simplicity, we fix the paper representation $\boldsymbol{x}_p$ to train a Parabel classifier $f_{\text{Parabel}}(\boldsymbol{x}_p) = \Pr(\boldsymbol{y}_p|\boldsymbol{x}_p)$.

During the prediction stage, given a paper p', the trained classifier predicts the probability that p' is relevant to each label l:

$$\hat{\boldsymbol{y}}_{p'} = f_{\text{Parabel}}(\boldsymbol{x}_{p'}). \tag{11}$$

Then, we can rank the labels according to $\hat{y}_{p',l}$ to predict a list of the most relevant labels. However, in practice, we find this strategy does not yield competitive prediction accuracy. This finding is consistent with the design in related studies [4, 31] that discrete lexical features should be considered together with continuous representations during classification. To add lexical features, motivated by [60], we propose a simple heuristic to re-rank the labels: Given a paper p', all labels whose name appears in the paper text $\mathcal{T}_{p'}$ will be ranked higher than those not appearing in $\mathcal{T}_{p'}$. This heuristic is equivalent to ranking the labels according to a modified score $\hat{z}_{p',l} = \hat{y}_{p',l} + \mathbf{1}(t_l \subseteq \mathcal{T}_{p'})$, where t_l stands for the textual name of label l.

4.3.2 Using Text + Metadata. When metadata information is available, we use OAG-BERT to jointly encode paper text and metadata (as illustrated in Figure 2(c)).

$$\widetilde{\boldsymbol{H}}_p = \text{OAG-BERT}(\mathcal{T}_p, \mathcal{M}_p). \tag{12}$$

The remaining steps exactly follow the text-only model. It is worth noting that since references are not involved during the pre-training of OAG-BERT, we can only use venues and authors as metadata here. Also, unlike the Transformer classifier that views each venue/author as one token, OAG-BERT tokenizes the textual name of venues/authors based on its vocabulary during encoding.

5 EXPERIMENTS

5.1 Setup

Datasets and Compared Methods. We have introduced the 20 datasets in Section 3. For each of the three classifiers, we test its performance when using text only, text+venue, text+author, and text+reference in order to check the effect of each metadata type separately. (Recall that OAG-BERT cannot take references as metadata features, so it does not have the text+reference variant.) For detailed hyperparameter settings, one can refer to Appendix A.2.

Evaluation Metrics. Following previous studies on multi-label text classification [21, 44, 59], we adopt P@k and NDCG@k as evaluation metrics, where $k = 1, 3, 5$. Given a paper p, let rank(i) be the index of the i-th highest predicted label according to each classifier, then

$$\begin{aligned} \text{P@}k &= \frac{1}{k}\sum_{i=1}^{k} y_{p,\text{rank}(i)}. \\ \text{DCG@}k = \sum_{i=1}^{k} \frac{y_{p,\text{rank}(i)}}{\log(i+1)}, \quad \text{NDCG@}k &= \frac{\text{DCG@}k}{\sum_{i=1}^{\min(k,\|\boldsymbol{y}_p\|_0)} \frac{1}{\log(i+1)}}. \end{aligned} \tag{13}$$

5.2 Overall Analysis

Table 2 shows the P@k and NDCG@k scores of the three classifiers on the 20 datasets. We run each experiment 5 times with the average score reported. We conduct two-tailed t-tests to check the statistical significance of metadata's effect. To be specific, given a field and a classifier, if a score is significantly improved (with p-value < 0.05) after using a certain type of metadata in comparison with using text only, we mark the score as blue in Table 2; if a score significantly deteriorates (with p-value < 0.05), we mark the score as red.

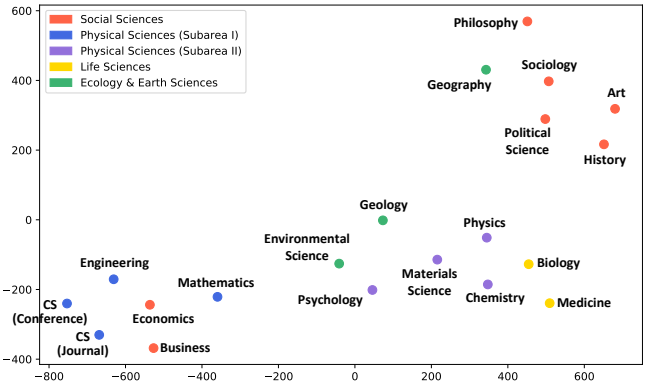


Figure 3: We represent each field with a 24-dimensional vector based on the effect of venue, author, and reference information on the three classifiers. Then, we apply t-SNE [27] to visualize these fields in a 2-dimensional space. The color scheme highlights several high-level scientific areas, following the major clusters of science detected by [35, 51] and suggesting similar effects of metadata within each area.

From Table 2, we can observe that: (1) The effect of metadata varies remarkably across different fields. For instance, author information is significantly helpful in Computer Science and Engineering when Parabel is utilized as the classifier but becomes harmful in Biology and Medicine when the same classifier is adopted. References are useful in Business and Economics when Parabel is employed but become disadvantageous in the same fields if Transformer is the tagger. (2) Venues are consistently beneficial to scientific literature tagging in almost all cases. To be specific, all 20 fields significantly benefit from venues[3] when Parabel is used, 18 fields when Transformer is used, and 18 fields when OAG-BERT is used. In the remaining 4 cases where venue information is not beneficial, neither is it harmful. (3) Authors are helpful in a majority of (15 and 17, respectively) fields when Parabel and OAG-BERT are the taggers but rarely help when Transformer is adopted. References are useful in even fewer cases. The overall performance of the three types of metadata is reflected by the macro average of P@k and NDCG@k scores over the 20 datasets, which are shown in Table 3. The reason why authors and references do not work in the Transformer classifier is that their embeddings need to be trained from scratch without good initialization. In contrast, leveraging venues only incurs a small number of additional parameters. Therefore, if one aims to utilize author and reference information, some embedding pre-training techniques [59] may help.

5.3 Effect in Different Fields

In this section, we examine the effect of metadata in different fields. Given a field F, to describe the effect of metadata in F, we construct a 24-dimensional vector in the following way: There are three classifiers (i.e., Parabel, Transformer, and OAG-BERT) and three types of metadata (i.e., Venue, Author, and Reference) studied

[3] We say one field *significantly benefits from* one type of metadata (when using a certain classifier) if at least one of the five metrics is significantly improved (with p-value < 0.05) after incorporating that type of metadata.

Table 2: P@k and NDCG@k scores on the 20 datasets. Blue : significantly better than using text only (p-value < 0.05). Red : significantly worse than using text only (p-value < 0.05). "–": OAG-BERT cannot take references as metadata signals.

Field	Input	Parabel [33]					Transformer [44]					OAG-BERT [24]				
		P@1	P@3	P@5	N@3	N@5	P@1	P@3	P@5	N@3	N@5	P@1	P@3	P@5	N@3	N@5
Art	Text	0.7203	0.4829	0.3392	0.7585	0.7921	0.6995	0.4396	0.3008	0.6956	0.7183	0.6982	0.4411	0.3030	0.7024	0.7253
	+Venue	0.7235	0.4861	0.3417	0.7630	0.7969	0.7109	0.4456	0.3038	0.7065	0.7279	0.7032	0.4476	0.3066	0.7116	0.7335
	+Author	0.7247	0.4827	0.3393	0.7598	0.7936	0.7060	0.4447	0.3035	0.7024	0.7238	0.7032	0.4454	0.3062	0.7085	0.7318
	+Reference	0.7186	0.4818	0.3389	0.7570	0.7913	0.7052	0.4432	0.3034	0.6997	0.7222	–	–	–	–	–
Philosophy	Text	0.8124	0.6033	0.4617	0.7578	0.7434	0.7493	0.5103	0.3712	0.6524	0.6219	0.7652	0.5548	0.4157	0.6980	0.6738
	+Venue	0.8164	0.6322	0.4909	0.7834	0.7738	0.7889	0.5725	0.4285	0.7159	0.6897	0.7679	0.5621	0.4271	0.7045	0.6852
	+Author	0.8128	0.6050	0.4637	0.7593	0.7455	0.7544	0.5151	0.3743	0.6577	0.6262	0.7673	0.5594	0.4206	0.7030	0.6799
	+Reference	0.8023	0.5968	0.4581	0.7478	0.7344	0.7442	0.5072	0.3698	0.6463	0.6160	–	–	–	–	–
Geography	Text	0.6233	0.4114	0.3015	0.6331	0.6640	0.5239	0.3140	0.2237	0.4887	0.5036	0.6220	0.3997	0.2843	0.6158	0.6341
	+Venue	0.6283	0.4147	0.3036	0.6384	0.6694	0.5457	0.3251	0.2304	0.5083	0.5223	0.6244	0.4036	0.2879	0.6213	0.6406
	+Author	0.6215	0.4113	0.3019	0.6321	0.6637	0.5241	0.3141	0.2239	0.4888	0.5038	0.6237	0.4046	0.2885	0.6219	0.6414
	+Reference	0.6135	0.4104	0.3021	0.6277	0.6603	0.4960	0.3030	0.2181	0.4665	0.4835	–	–	–	–	–
Business	Text	0.6601	0.4775	0.3666	0.6192	0.6351	0.6375	0.4344	0.3238	0.5736	0.5780	0.6687	0.4859	0.3648	0.6328	0.6408
	+Venue	0.6667	0.4811	0.3690	0.6247	0.6402	0.6550	0.4435	0.3293	0.5878	0.5914	0.6693	0.4899	0.3688	0.6371	0.6462
	+Author	0.6626	0.4782	0.3673	0.6208	0.6367	0.6265	0.4255	0.3161	0.5619	0.5649	0.6701	0.4886	0.3679	0.6362	0.6455
	+Reference	0.6587	0.4798	0.3694	0.6219	0.6389	0.5872	0.3985	0.2947	0.5237	0.5242	–	–	–	–	–
Sociology	Text	0.6173	0.4207	0.3057	0.6219	0.6509	0.5851	0.3611	0.2514	0.5489	0.5614	0.6177	0.4004	0.2799	0.5998	0.6144
	+Venue	0.6227	0.4257	0.3080	0.6297	0.6573	0.6026	0.3704	0.2576	0.5644	0.5768	0.6195	0.4046	0.2834	0.6048	0.6202
	+Author	0.6182	0.4206	0.3054	0.6220	0.6505	0.5970	0.3625	0.2517	0.5548	0.5664	0.6191	0.4027	0.2820	0.6026	0.6179
	+Reference	0.6112	0.4210	0.3068	0.6200	0.6500	0.5805	0.3573	0.2491	0.5395	0.5516	–	–	–	–	–
History	Text	0.7164	0.4574	0.3173	0.7480	0.7799	0.6941	0.4210	0.2869	0.6928	0.7174	0.6591	0.4122	0.2822	0.6776	0.7029
	+Venue	0.7199	0.4615	0.3200	0.7537	0.7856	0.7008	0.4264	0.2897	0.7010	0.7245	0.6602	0.4147	0.2841	0.6809	0.7064
	+Author	0.7144	0.4571	0.3171	0.7469	0.7789	0.6988	0.4233	0.2881	0.6959	0.7198	0.6627	0.4161	0.2853	0.6832	0.7089
	+Reference	0.7142	0.4565	0.3169	0.7463	0.7786	0.6999	0.4248	0.2892	0.6990	0.7233	–	–	–	–	–
Political Science	Text	0.7723	0.5153	0.3767	0.7252	0.7291	0.7624	0.4863	0.3429	0.6892	0.6797	0.7327	0.4805	0.3393	0.6721	0.6624
	+Venue	0.7776	0.5216	0.3817	0.7327	0.7374	0.7723	0.4948	0.3494	0.7004	0.6910	0.7320	0.4843	0.3438	0.6756	0.6672
	+Author	0.7737	0.5173	0.3779	0.7275	0.7313	0.7644	0.4882	0.3439	0.6913	0.6810	0.7356	0.4850	0.3436	0.6771	0.6680
	+Reference	0.7633	0.5107	0.3750	0.7182	0.7237	0.7549	0.4812	0.3392	0.6811	0.6712	–	–	–	–	–
Environmental Science	Text	0.6484	0.4254	0.3087	0.6990	0.7299	0.6447	0.4003	0.2856	0.6706	0.6964	0.7185	0.4465	0.3156	0.7449	0.7663
	+Venue	0.6488	0.4269	0.3094	0.7005	0.7313	0.6469	0.4023	0.2867	0.6730	0.6981	0.7196	0.4468	0.3157	0.7454	0.7669
	+Author	0.6484	0.4267	0.3095	0.7002	0.7313	0.6333	0.3945	0.2816	0.6590	0.6845	0.7189	0.4474	0.3166	0.7461	0.7683
	+Reference	0.6486	0.4278	0.3112	0.7016	0.7340	0.6126	0.3843	0.2752	0.6388	0.6648	–	–	–	–	–
Economics	Text	0.7288	0.5842	0.4793	0.6483	0.6301	0.7375	0.5785	0.4635	0.6448	0.6167	0.6861	0.5651	0.4612	0.6229	0.6037
	+Venue	0.7339	0.5885	0.4830	0.6534	0.6353	0.7460	0.5848	0.4690	0.6521	0.6243	0.6868	0.5677	0.4645	0.6251	0.6069
	+Author	0.7288	0.5856	0.4809	0.6496	0.6315	0.7224	0.5607	0.4477	0.6258	0.5967	0.6875	0.5676	0.4655	0.6254	0.6080
	+Reference	0.7306	0.5866	0.4826	0.6511	0.6340	0.6867	0.5303	0.4203	0.5921	0.5609	–	–	–	–	–
Computer Science (Conference)	Text	0.7158	0.5594	0.4597	0.6222	0.5909	0.6980	0.5287	0.4196	0.5924	0.5498	0.6516	0.5299	0.4349	0.5835	0.5541
	+Venue	0.7196	0.5638	0.4638	0.6268	0.5958	0.7079	0.5400	0.4300	0.6040	0.5618	0.6508	0.5308	0.4364	0.5832	0.5541
	+Author	0.7164	0.5608	0.4610	0.6234	0.5923	0.6770	0.5122	0.4068	0.5727	0.5308	0.6519	0.5315	0.4375	0.5847	0.5561
	+Reference	0.7151	0.5614	0.4618	0.6236	0.5927	0.6377	0.4777	0.3774	0.5353	0.4940	–	–	–	–	–
Engineering	Text	0.7881	0.6383	0.5274	0.7064	0.6776	0.7801	0.6154	0.4961	0.6852	0.6461	0.6780	0.5668	0.4757	0.6234	0.6058
	+Venue	0.7904	0.6399	0.5285	0.7087	0.6798	0.7818	0.6168	0.4980	0.6869	0.6486	0.6819	0.5699	0.4787	0.6270	0.6101
	+Author	0.7885	0.6397	0.5286	0.7078	0.6790	0.7679	0.5995	0.4820	0.6686	0.6286	0.6784	0.5686	0.4788	0.6251	0.6091
	+Reference	0.7879	0.6398	0.5288	0.7079	0.6793	0.7261	0.5579	0.4453	0.6242	0.5832	–	–	–	–	–
Psychology	Text	0.7784	0.6272	0.5133	0.7073	0.6889	0.8035	0.6434	0.5172	0.7274	0.7007	0.7503	0.6148	0.4986	0.6909	0.6686
	+Venue	0.7871	0.6323	0.5173	0.7143	0.6958	0.8108	0.6481	0.5214	0.7339	0.7078	0.7536	0.6200	0.5047	0.6964	0.6758
	+Author	0.7775	0.6283	0.5154	0.7078	0.6903	0.7918	0.6300	0.5062	0.7126	0.6858	0.7489	0.6165	0.5015	0.6921	0.6712
	+Reference	0.7821	0.6334	0.5209	0.7136	0.6973	0.7776	0.6205	0.4979	0.7017	0.6749	–	–	–	–	–
Computer Science (Journal)	Text	0.7075	0.5611	0.4645	0.6209	0.5862	0.7125	0.5477	0.4415	0.6110	0.5663	0.6356	0.5266	0.4416	0.5760	0.5481
	+Venue	0.7098	0.5639	0.4673	0.6238	0.5893	0.7146	0.5501	0.4439	0.6134	0.5689	0.6359	0.5296	0.4465	0.5789	0.5529
	+Author	0.7094	0.5629	0.4662	0.6228	0.5883	0.6937	0.5278	0.4241	0.5897	0.5449	0.6362	0.5299	0.4467	0.5792	0.5533
	+Reference	0.7089	0.5636	0.4675	0.6234	0.5894	0.6534	0.4920	0.3936	0.5503	0.5062	–	–	–	–	–
Geology	Text	0.8618	0.7327	0.6213	0.8052	0.7765	0.8752	0.7412	0.6254	0.8147	0.7825	0.7832	0.6622	0.5605	0.7276	0.7004
	+Venue	0.8634	0.7347	0.6234	0.8071	0.7787	0.8779	0.7455	0.6299	0.8190	0.7875	0.7851	0.6637	0.5631	0.7295	0.7031
	+Author	0.8601	0.7319	0.6215	0.8036	0.7756	0.8641	0.7310	0.6171	0.8018	0.7693	0.7814	0.6600	0.5606	0.7251	0.6990
	+Reference	0.8432	0.7197	0.6138	0.7886	0.7622	0.8485	0.7162	0.6031	0.7848	0.7503	–	–	–	–	–
Mathematics	Text	0.7406	0.6191	0.5326	0.6584	0.6192	0.7715	0.6368	0.5403	0.6796	0.6335	0.6174	0.5315	0.4641	0.5612	0.5334
	+Venue	0.7421	0.6210	0.5341	0.6603	0.6211	0.7756	0.6421	0.5455	0.6848	0.6390	0.6187	0.5351	0.4689	0.5643	0.5376
	+Author	0.7408	0.6214	0.5346	0.6603	0.6212	0.7642	0.6294	0.5330	0.6718	0.6252	0.6175	0.5347	0.4687	0.5637	0.5372
	+Reference	0.7412	0.6237	0.5383	0.6623	0.6242	0.7324	0.5974	0.4999	0.6390	0.5895	–	–	–	–	–
Materials Science	Text	0.8193	0.6638	0.5436	0.7746	0.7720	0.8670	0.6994	0.5696	0.8203	0.8157	0.7490	0.5935	0.4808	0.6932	0.6845
	+Venue	0.8233	0.6667	0.5459	0.7786	0.7761	0.8692	0.7009	0.5707	0.8228	0.8180	0.7496	0.5943	0.4823	0.6946	0.6867
	+Author	0.8203	0.6659	0.5459	0.7766	0.7745	0.8558	0.6879	0.5597	0.8069	0.8016	0.7469	0.5928	0.4812	0.6921	0.6843
	+Reference	0.8206	0.6654	0.5462	0.7762	0.7747	0.8510	0.6840	0.5567	0.8021	0.7970	–	–	–	–	–
Physics	Text	0.8367	0.7228	0.6219	0.7597	0.7050	0.8739	0.7552	0.6490	0.7941	0.7368	0.7461	0.6399	0.5585	0.6730	0.6300
	+Venue	0.8452	0.7269	0.6248	0.7650	0.7095	0.8787	0.7593	0.6520	0.7985	0.7407	0.7491	0.6419	0.5609	0.6754	0.6328
	+Author	0.8372	0.7235	0.6236	0.7602	0.7061	0.8665	0.7472	0.6406	0.7857	0.7276	0.7415	0.6372	0.5566	0.6698	0.6270
	+Reference	0.8340	0.7197	0.6213	0.7563	0.7029	0.8682	0.7486	0.6423	0.7873	0.7294	–	–	–	–	–
Biology	Text	0.8446	0.7503	0.6609	0.7804	0.7292	0.8764	0.7743	0.6798	0.8069	0.7524	0.7448	0.6517	0.5772	0.6804	0.6371
	+Venue	0.8474	0.7523	0.6626	0.7829	0.7317	0.8791	0.7766	0.6813	0.8094	0.7544	0.7484	0.6571	0.5840	0.6856	0.6436
	+Author	0.8433	0.7495	0.6610	0.7793	0.7286	0.8652	0.7574	0.6600	0.7909	0.7332	0.7427	0.6528	0.5806	0.6808	0.6393
	+Reference	0.8456	0.7505	0.6618	0.7807	0.7300	0.8722	0.7668	0.6697	0.8000	0.7432	–	–	–	–	–
Chemistry	Text	0.8448	0.7208	0.6211	0.7720	0.7310	0.8866	0.7658	0.6587	0.8187	0.7749	0.7480	0.6311	0.5334	0.6768	0.6312
	+Venue	0.8485	0.7229	0.6229	0.7756	0.7347	0.8896	0.7668	0.6590	0.8214	0.7774	0.7503	0.6361	0.5413	0.6819	0.6390
	+Author	0.8444	0.7220	0.6232	0.7727	0.7324	0.8811	0.7577	0.6499	0.8105	0.7654	0.7480	0.6346	0.5403	0.6794	0.6366
	+Reference	0.8425	0.7182	0.6204	0.7694	0.7295	0.8810	0.7568	0.6484	0.8102	0.7649	–	–	–	–	–
Medicine	Text	0.7894	0.6491	0.5325	0.7335	0.7227	0.8421	0.6935	0.5669	0.7849	0.7711	0.6935	0.5744	0.4789	0.6458	0.6413
	+Venue	0.7936	0.6528	0.5360	0.7379	0.7275	0.8436	0.6951	0.5685	0.7867	0.7729	0.7012	0.5818	0.4884	0.6534	0.6516
	+Author	0.7809	0.6428	0.5297	0.7246	0.7148	0.8291	0.6800	0.5565	0.7688	0.7549	0.6873	0.5727	0.4804	0.6424	0.6397
	+Reference	0.7913	0.6520	0.5362	0.7364	0.7264	0.8394	0.6902	0.5633	0.7817	0.7672	–	–	–	–	–

Table 3: Macro average of P@k and NDCG@k over the 20 datasets. Blue, Red, and "–": the same meaning as in Table 2.

	Input	Parabel [33]					Transformer [44]					OAG-BERT [24]				
		P@1	P@3	P@5	N@3	N@5	P@1	P@3	P@5	N@3	N@5	P@1	P@3	P@5	N@3	N@5
Macro Average	Text	0.7513	0.5811	0.4678	0.7076	0.6977	0.7510	0.5673	0.4507	0.6896	0.6712	0.6983	0.5354	0.4275	0.6549	0.6429
	+Venue	0.7554	0.5858	0.4717	0.7130	0.7033	0.7599	0.5753	0.4572	0.6995	0.6812	0.7004	0.5391	0.4318	0.6588	0.6480
	+Author	0.7512	0.5817	0.4687	0.7079	0.6983	0.7442	0.5594	0.4433	0.6809	0.6617	0.6984	0.5374	0.4305	0.6569	0.6461
	+Reference	0.7487	0.5809	0.4689	0.7065	0.6977	0.7277	0.5469	0.4328	0.6652	0.6459	–	–	–	–	–

in our experiments, so there are $3 \times 3 - 1 = 8$ (classifier, metadata) combinations, since (OAG-BERT, Reference) is not applicable. For each of the 8 (classifier, metadata) combinations, we calculate the relative performance change of P@1, P@3, and P@5 by comparing the classifier using text only and the classifier using text together with the metadata. For example, given the (Parabel, Venue) combination, we calculate the following 3 values:

$$\frac{\text{P@}k\,(\text{Parabel, Text + Venue}) - \text{P@}k\,(\text{Parabel, Text})}{\text{P@}k\,(\text{Parabel, Text})}, \quad k = 1, 3, 5. \tag{14}$$

In total, we will have $3 \times 8 = 24$ values, which can form a 24-dimensional vector $\boldsymbol{u}_F$. We compute $\boldsymbol{u}_F$ for all 20 fields and then apply t-SNE [27] to visualize these vectors in a 2-dimensional space.

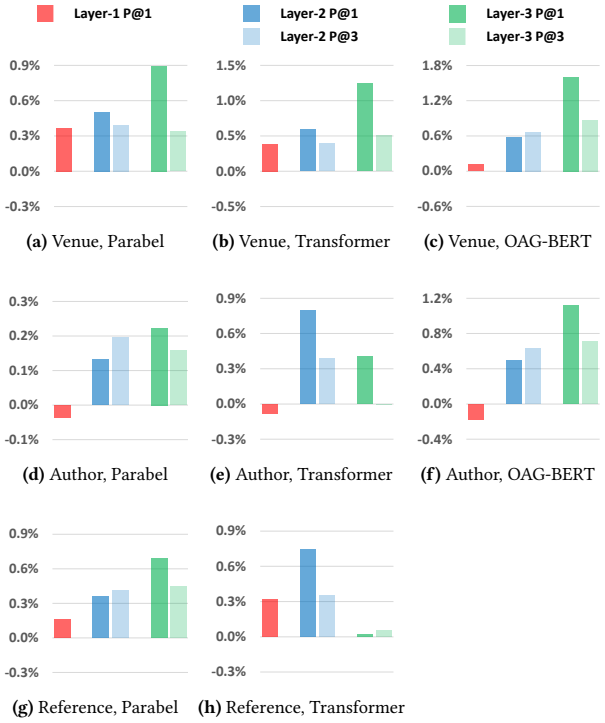
The visualization result is shown in Figure 3, where we color each field according to high-level scientific areas/subareas detected by [35, 51]. To be specific, Rosvall and Bergstrom [35] cluster scientific fields into four major areas – social sciences, physical sciences, life sciences, and ecology & earth sciences – based on a large-scale paper citation network. The cluster of physical sciences is further split into two subareas, one of which is related to mathematics and computer science, and the other to physics and chemistry; Yin et al. [51] observe a similar partitioning of the 19 fields when studying public uses of scientific papers in each field.

In Figure 3, we find that fields with the same color are often embedded closer, indicating that metadata have similar effects in fields belonging to the same area/subarea. This finding is very useful when we need to extrapolate the experience of using metadata in one field to a similar field. For example, one may have known that references are beneficial and authors are harmful when using Parabel in the Medicine field. Based on our finding, the same pattern can be deduced when using Parabel in the Biology field because Medicine and Biology are both life sciences. According to Table 2, this deduction is correct.

Meanwhile, we observe two exceptions: (1) The area of social sciences is split into two clusters, one of which contains Art, History, Philosophy, Sociology, and Political Science, and the other contains Economics and Business. The effects of metadata in the Business and Economics fields are more similar to those in Mathematics and Computer Science. This is possibly because a large proportion of Economics and Business papers rely on quantitative analysis of massive data to draw conclusions, which is different from the paradigm in most Art, History, and Philosophy papers. (2) Geography is embedded closer to social sciences than it is to Geology and Environmental Science. This is possibly because Geography is an interdisciplinary field. The physical geography subfield is more related to earth sciences while the human geography subfield is closer to social sciences. Overall, we still observe commonalities in the effects of metadata within high-level areas/subareas.

5.4 Effect at Different Label Granularities

Now we examine the effect of metadata when predicting labels at different granularity levels. The MAG taxonomy [37] has 6 layers (from the most coarse-grained Layer 0 to the most fine-grained Layer 5). Layer-0 labels are the 19 fields excluded from our label space. Across all 20 datasets, 85.2%, 86.5%, 70.4%, 35.0%, and 13.2% of the papers have ground-truth Layer-1, Layer-2, Layer-3, Layer-4, and Layer-5 labels, respectively. Due to the low proportions of papers related to Layer-4 and Layer-5 tags, we only study the effect of metadata on predicting Layer-1, Layer-2, and Layer-3 tags.



(a) Venue, Parabel **(b)** Venue, Transformer **(c)** Venue, OAG-BERT

(d) Author, Parabel **(e)** Author, Transformer **(f)** Author, OAG-BERT

(g) Reference, Parabel **(h)** Reference, Transformer

Figure 4: The effect of metadata on Layer-j P@k scores (averaged over the fields that significantly benefit from leveraging that type of metadata using the classifier).

Table 4: The fields that benefit the most from leveraging venue information in terms of Layer-1 P@1, Layer-2 P@1, and Layer-3 P@1. "T": text only. "+V": text+venue. "Δ": absolute performance improvement.

		L1 P@1			L2 P@1			L3 P@1
Parabel [33]								
Physics	T	0.7178	Philosophy	T	0.7846	Philosophy	T	0.6118
	+V	0.7284		+V	0.8170		+V	0.7209
	Δ	1.06%		Δ	3.24%		Δ	10.91%
Transformer [44]								
Physics	T	0.7393	Philosophy	T	0.6728	Philosophy	T	0.4105
	+V	0.7495		+V	0.7504		+V	0.6071
	Δ	1.02%		Δ	7.76%		Δ	19.66%
OAG-BERT [24]								
Geography	T	0.6626	Art	T	0.6747	History	T	0.3455
	+V	0.6679		+V	0.6879		+V	0.4130
	Δ	0.53%		Δ	1.32%		Δ	6.75%

Given one classifier and one type of metadata, we consider all fields that significantly benefit from the metadata using the classifier. In these fields, we calculate the Layer-j P@k scores ($(j,k) = (1,1),(2,1),(2,3),(3,1),(3,3)$). The definition of Layer-j P@k is very similar to that of P@k except that Layer-j P@k considers the k most confident labels at Layer j instead of in the whole label space. When calculating Layer-j P@k, we focus on papers with at least one ground-truth Layer-j label. Then, we compute the absolute performance change of Layer-j P@k scores after incorporating the considered type of metadata. The results are shown in Figure 4.

From Figure 4, we find that: (1) On average, venues can help scientific literature tagging for not only coarse-grained tags but also fine-grained ones. This may be counterintuitive from computer scientists' perspective because CS venues (e.g., "*WWW*") can hardly indicate very fine tags (e.g., "Link Farm"). Indeed, in the Computer Science (Conference) dataset, the contribution of venues on Layer-3 P@1 is subtle (e.g., 0.19% when using Parabel, -0.11% when using Transformer). However, in fields other than Computer Science, some venues do carry very fine-grained signals. For example, there are two venues "*Journal of Roman Archaeology*" and "*Mediaeval Studies*" in History and Philosophy, respectively. These two venues may strongly imply a paper's relevance to "Classical Archaeology" and "Medievalism", which are Layer-2 and Layer-3 tags, respectively. In Table 4, we list the fields that benefit the most from leveraging venue information in terms of Layer-1 P@1, Layer-2 P@1, and Layer-3 P@1, where we do observe that Philosophy and History are the biggest beneficiaries in terms of Layer-3 P@1. (2) Different from venues, authors are beneficial to fine-grained tagging but harmful to coarse-grained prediction. This observation consistently holds across all three classifiers.

6 RELATED WORK

Extreme Multi-Label Text Classification. In keeping with our discussion in Section 4, we divide related studies on extreme multi-label classification into three major categories. (1) *Bag-of-words classifiers* [2, 16, 19, 31–34, 39, 49, 50, 54] take sparse tf–idf features as input. To improve model efficiency, 1-vs-all approaches such as DiSMEC [2] and PPDSparse [49] explore parallelism and model size reduction via model weight truncation. In another direction, tree-based approaches apply various partitioning techniques on the large label space. For example, Parabel [33] partitions the labels to a balanced tree structure using 2-means clustering; Bonsai [19] improves Parabel by allowing multi-way and unbalanced partitions; XR-Linear [54] improves Parabel by incorporating various hard negative sampling schemes; AnnexML [39] partitions the labels via graph-based nearest neighbor indices. (2) *Sequence-based classifiers* [21, 44, 45, 47, 53, 59] employ deep neural architectures such as CNNs (e.g., XML-CNN [21]), RNNs (e.g., MeshProbeNet [45] and AttentionXML [53]), and Transformers (e.g., BertXML [44]) to learn semantic representations of input text sequences for classification. There are also studies aggregating shallow word embeddings and/or applying MLP layers to obtain document embeddings, such as Slice [15], DeepXML [8], DECAF [29], GalaXC [36], and ECLARE [30]. (3) *Pre-trained language model classifiers* [4, 17, 48, 52, 56, 60] propose to transfer the knowledge learned by PLMs from web-scale corpora to the classification task. For example, X-Transformer [4] complements the output of BERT [10], XLNet [46], or RoBERTa [25] with sparse tf–idf features; LightXML [17] adopts PLMs as the text encoder and performs label shortlist and re-ranking with the same PLM; XR-Transformer [56] further proposes fast multi-resolution PLM fine-tuning. Despite the success of these models in extreme multi-label classification, they mainly focus on classifying plain text sequences and are less aware of document metadata. In contrast, our work proposes several straightforward ways to enhance text classifiers with metadata signals.

Scientific Literature Tagging. Classifying academic papers is a common evaluation task in text mining [6, 28, 58] and graph mining [11, 14] studies. However, most studies consider coarse-grained paper classification only (e.g., with 5-20 categories in the label space), the result of which is not subdivided enough to satisfy users' fine-grained interests. To tag papers on PubMed with fine-grained medical subject headings, the task of MeSH indexing has been extensively studied [9, 22, 31, 45, 52]. However, these models only use text or text+venue as input, leaving the effect of authors and references unexplored. Recently, Zhang et al. [59, 60] and Ye et al. [47] make use of metadata to tag papers with fields of study in MAG. Still, these studies are restricted to computer science and biomedicine fields only. In comparison, our work conducts a systematic study across 19 fields and three major types of classifiers.

7 CONCLUSIONS AND FUTURE WORK

In this work, we examine the effect of metadata on scientific literature tagging in 19 fields using three classifiers. Our results provide the following insights to practitioners aiming at building accurate scientific literature taggers: First, while previous metadata-aware approaches often directly use all types of available metadata, we show that not all of them are always beneficial. It is important to select useful metadata features based on the classifier's type, the field, and the granularity level of predicted tags. Second, although the state-of-the-art models for text sequence modeling have been dominated by Transformer-based models, we demonstrate that simple bag-of-words classifiers work comparably well or even better in many cases for large-scale fine-grained paper tagging, and may be more effective in leveraging different types of metadata. Third, despite the varying effects of different metadata types in different fields, the gain or loss induced by each metadata type is rather consistent across similar fields.

This study explores each type of metadata separately to avoid confounders. However, different types of metadata (e.g., a venue and an author) may interact with each other and provide additional hints for classification. In the future, it is of our interest to study the composite effect of multiple types of metadata.

ACKNOWLEDGMENTS

We thank anonymous reviewers for their valuable and insightful feedback. Research was supported in part by the IBM-Illinois Discovery Accelerator Institute, US DARPA KAIROS Program No. FA8750-19-2-1004 and INCAS Program No. HR001121C0165, 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. 2019897, and the Institute for Geospatial Understanding through an Integrative Discovery Environment (I-GUIDE) by NSF under Award No. 2118329. 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.

A APPENDIX

A.1 Datasets

Supplementary to Table 1, Table 5 summarizes more dataset statistics in MAPLE, including the average numbers of authors, references, and labels per paper as well as the numbers of training, validation, and testing papers.

Table 5: More statistics of the 20 datasets in MAPLE.

Field	#Authors/ Paper	#References/ Paper	#Labels/ Paper	#Train Papers	#Valid Papers	#Test Papers
Art	1.314	2.813	2.425	39,901	9,975	8,497
Philosophy	1.117	8.121	3.715	40,696	10,174	8,426
Geography	3.625	31.793	2.231	44,710	11,178	17,995
Business	2.346	37.635	3.556	49,481	12,370	23,007
Sociology	1.592	24.474	2.318	60,815	15,204	14,189
History	1.218	4.405	1.938	80,830	20,208	12,109
Political Science	1.496	10.198	3.107	73,650	18,413	23,228
Environmental Science	4.301	34.156	2.138	70,479	17,620	35,846
Economics	1.966	26.608	4.961	119,309	29,827	29,534
Engineering	3.055	19.703	4.987	179,804	44,951	45,251
Psychology	3.805	41.201	5.032	245,288	61,322	66,344
CS (Conference)	3.389	17.417	6.089	136,322	34,081	92,990
CS (Journal)	3.296	21.968	5.976	214,616	53,654	142,333
Geology	3.552	36.510	5.957	284,022	71,006	76,806
Mathematics	2.151	19.068	6.433	338,454	84,613	67,484
Materials Science	4.829	26.838	4,428	783,289	195,822	358,620
Physics	9.841	23.509	7.271	905,613	226,403	237,967
Biology	5.611	36.137	8.306	1,125,605	281,401	181,772
Chemistry	4.293	28.382	6.288	1,252,531	313,133	284,292
Medicine	5.647	12.888	5.376	1,620,165	405,041	620,899

A.2 Hyperparameters

In each of the 20 datasets, we remove metadata instances that appear in less than 5 papers. We adopt the same hyperparameter configuration when using text only, text+venue, text+author, and text+reference as input. The code source and hyperparameters of each classifier are introduced below.

A.2.1 Parabel. [4] We remove words appearing in less than 5 papers. All parameters are set by default. Specifically, the number of threads $T = 1$; the number of trees $t = 3$; the maximum number of labels in a leaf node $m = 100$; the beam search width in prediction $B = 10$.

A.2.2 Transformer. [5] The number of Transformer layers is 1; the number of attention heads is 2; the number of [CLS] tokens $C = 10$; the embedding dimension is 100; the maximum sequence length is 500; the batch size is 256. We adopt GloVe.6B.100d as initialized word embeddings. For (the number of training epochs, the number of warm-up epochs), we use (100, 20) for Art, Philosophy, and Geography, (80, 16) for Business, Sociology, History, and Political Science, (60, 12) for Environmental Science, Economics, Engineering, and Computer Science, (40, 8) for Psychology, Geology, and Mathematics, (20, 4) for Materials Science, Physics, and Biology, and (15, 3) for Chemistry and Medicine. Other hyperparameters are set by default.

A.2.3 OAG-BERT. [6] We use "oagbert-v2" as the PLM. After PLM encoding, we fix paper embeddings to train a Parabel classifier. All parameters of Parabel are set by default.

[4] http://manikvarma.org/code/Parabel/download.html
[5] https://github.com/XunGuangxu/CorNet (We use the BertXML classifier in this GitHub repository. Although the model is called BertXML, it trains a Transformer architecture from scratch without BERT initialization.)
[6] https://github.com/THUDM/OAG-BERT

A.3 More on the Effect of Metadata at Different Label Granularities

Supplementary to Table 4, Table 6 shows the fields that benefit the most from leveraging author or reference information in terms of Layer-1 P@1, Layer-2 P@1, and Layer-3 P@1. We find that authors and references are beneficial to the Art, Philosophy, and History fields in many cases. The possible reason is that each paper in these fields has a small number of authors and references (according to the statistics in Table 5). Therefore, the author or reference list may contain fewer confounding signals and be more topic-indicative.

Table 6: The fields that benefit the most from leveraging author or reference information in terms of Layer-1 P@1, Layer-2 P@1, and Layer-3 P@1. "T": text only. "+A": text+author. "+R": text+reference. "Δ": absolute performance improvement.

		L1 P@1			L2 P@1			L3 P@1
Parabel [33]								
Mathematics	T	0.5960	Political Science	T	0.7928	Philosophy	T	0.6118
	+A	0.5971		+A	0.7969		+A	0.6160
	Δ	0.11%		Δ	0.41%		Δ	0.42%
Psychology	T	0.6992	Business	T	0.6704	Business	T	0.6124
	+R	0.7086		+R	0.6801		+R	0.6265
	Δ	0.94%		Δ	0.97%		Δ	1.41%
Transformer [44]								
History	T	0.5471	Philosophy	T	0.6728	Art	T	0.4859
	+A	0.5489		+A	0.6900		+A	0.4910
	Δ	0.18%		Δ	1.72%		Δ	0.51%
History	T	0.5471	Art	T	0.7859	History	T	0.4857
	+R	0.5537		+R	0.7953		+R	0.4890
	Δ	0.66%		Δ	0.94%		Δ	0.33%
OAG-BERT [24]								
Art	T	0.6241	Art	T	0.6747	History	T	0.3455
	+A	0.6254		+A	0.6875		+A	0.3819
	Δ	0.13%		Δ	1.28%		Δ	3.64%

A.4 Effect of Metadata on Efficiency

Now we report the effect of metadata on model efficiency. Table 7 shows the average relative training time increase of the three classifiers across the 20 datasets after incorporating venues, authors, and references, respectively. All models are run on Intel Xeon E5-2680 v2 @ 2.80GHz and one NVIDIA GeForce GTX 1080 Ti GPU (if a GPU is needed). We can observe a significant increase in training time after adding references as features. When Parabel is the classifier, the increase is due to a much longer bag-of-words vector used to represent a paper; when Transformer is the classifier, the increase is caused by a large number of additional parameters (i.e., reference embeddings), which make the model converge more slowly.

Table 7: Average relative increase in training time across 20 datasets after incorporating one type of metadata.

	Parabel [33]	Transformer [44]	OAG-BERT [24]
+Venue	+0.15%	+0.13%	+0.11%
+Author	+0.72%	+1.86%	+0.11%
+Reference	+22.68%	+10.75%	–

Figure 5 shows the training time of the three classifiers on the 20 datasets. The reported training time is an average over 20 runs (i.e., 5 runs × {text only, text+venue, text+author, text+reference}) when Parabel and Transformer are tested, or 15 runs (i.e., 5 runs × {text only, text+venue, text+author}) when OAG-BERT is tested. Among the three classifiers, Parabel is consistently the most efficient across the 20 datasets; Transformer spends more time than OAG-BERT on

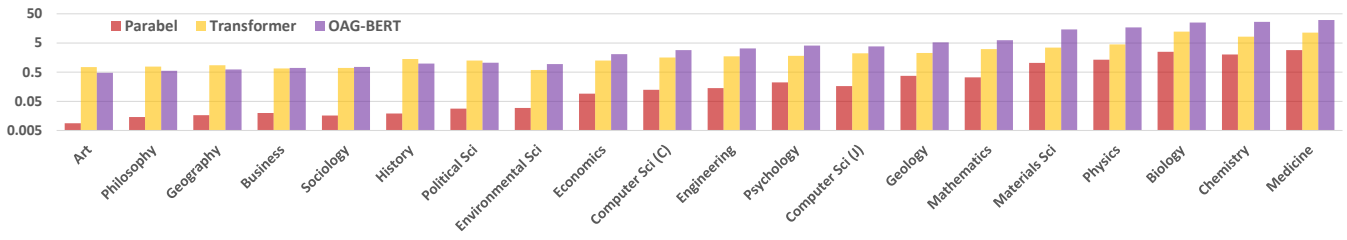


Figure 5: Training time (in hours) of the three classifiers on the 20 datasets.

Table 8: Statistics of the three additional datasets with MeSH labels.

Field	Paper Source	#Papers	#Labels	#Venues	#Authors	#References	#Authors/ Paper	#References/ Paper	#Labels/ Paper	#Train Papers	#Valid Papers	#Test Papers
Biology-MeSH	Journal	1,379,393	25,039	100	2,486,814	6,876,739	5.686	40.043	13.870	985,364	246,341	147,688
Chemistry-MeSH	Journal	762,129	21,585	87	1,498,358	5,928,908	4.741	34.344	10.984	511,814	127,954	122,361
Medicine-MeSH	Journal	1,536,660	25,188	100	2,791,165	7,190,021	5.254	20.931	11.819	1,020,969	255,242	260,449

Table 9: P@k and NDCG@k scores on the three additional datasets. Blue, Red, and "–": the same meaning as in Table 2.

Field	Input	Parabel [33]					Transformer [44]					OAG-BERT [24]				
		P@1	P@3	P@5	N@3	N@5	P@1	P@3	P@5	N@3	N@5	P@1	P@3	P@5	N@3	N@5
Biology-MeSH	Text	0.8924	0.7964	0.7088	0.8194	0.7576	0.9112	0.8098	0.7193	0.8341	0.7698	0.7506	0.6185	0.5460	0.6490	0.5945
	+Venue	0.8934	0.7976	0.7101	0.8206	0.7587	0.9119	0.8105	0.7194	0.8348	0.7701	0.7520	0.6200	0.5473	0.6504	0.5958
	+Author	0.8932	0.7985	0.7112	0.8212	0.7596	0.9090	0.8028	0.7093	0.8281	0.7614	0.7504	0.6184	0.5464	0.6488	0.5946
	+Reference	0.8976	0.8058	0.7198	0.8279	0.7674	0.9079	0.7988	0.7034	0.8248	0.7565	–	–	–	–	–
Chemistry-MeSH	Text	0.8447	0.7340	0.6407	0.7603	0.6947	0.8445	0.7113	0.6082	0.7427	0.6684	0.6971	0.5804	0.5099	0.6071	0.5557
	+Venue	0.8453	0.7354	0.6421	0.7615	0.6960	0.8465	0.7151	0.6121	0.7460	0.6720	0.6995	0.5826	0.5132	0.6093	0.5587
	+Author	0.8450	0.7350	0.6419	0.7611	0.6957	0.8439	0.7105	0.6066	0.7419	0.6670	0.6981	0.5819	0.5126	0.6086	0.5580
	+Reference	0.8490	0.7409	0.6491	0.7667	0.7023	0.8248	0.6802	0.5743	0.7139	0.6366	–	–	–	–	–
Medicine-MeSH	Text	0.9673	0.8410	0.7454	0.8712	0.8075	0.9683	0.8567	0.7583	0.8842	0.8191	0.7343	0.6229	0.5629	0.6491	0.6089
	+Venue	0.9673	0.8422	0.7472	0.8722	0.8090	0.9688	0.8585	0.7606	0.8856	0.8210	0.7370	0.6235	0.5607	0.6503	0.6081
	+Author	0.9671	0.8342	0.7424	0.8656	0.8039	0.9687	0.8495	0.7496	0.8786	0.8117	0.7347	0.6185	0.5549	0.6459	0.6027
	+Reference	0.9666	0.8382	0.7463	0.8687	0.8073	0.9662	0.8470	0.7458	0.8762	0.8084	–	–	–	–	–

small datasets, but the situation is reversed on moderate-sized and large datasets.

A.5 Additional Datasets with MeSH Labels

To further strengthen our findings, we construct three datasets with Medical Subject Headings (MeSH) terms [7] as their labels, which are curated by experts from the National Library of Medicine. To be specific, we take the three datasets, Biology, Chemistry, and Medicine, from MAPLE and obtain the ground-truth MeSH labels of each paper[7]. After removing papers not having MeSH labels, we get three new datasets, Biology-MeSH, Chemistry-MeSH, and Medicine-MeSH. Their statistics are shown in Table 8.

We run Parabel, Transformer, and OAG-BERT on the three new datasets. When running Transformer, for (the number of training epochs, the number of warm-up epochs), we use (20, 4) for all three datasets. When running OAG-BERT, we need to rerank those labels appearing in the paper text higher than those not. Note that a MeSH label may have multiple label names (i.e., one canonical name and 0, 1, or several entry terms, see the MeSH label "COVID-19"[8] as an example). Given a MeSH label, if any of its label names appears in the paper text, we view it as occurring in the paper. All other hyperparameters are the same as in Appendix A.2.

The P@k and NDCG@k scores of Parabel, Transformer, and OAG-BERT on the three new datasets are demonstrated in Table 9. In general, we still find that venues are beneficial to scientific literature tagging in almost all cases, while the effect of authors and references depends on the classifier's type and the field.

The three additional datasets are also available in our MAPLE benchmark: https://doi.org/10.5281/zenodo.7611544.

[7] https://learn.microsoft.com/en-us/academic-services/graph/reference-data-schema#paper-mesh

[8] https://meshb.nlm.nih.gov/record/ui?ui=D000086382

REFERENCES

[1] Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, et al. 2018. Construction of the Literature Graph in Semantic Scholar. In *NAACL-HLT'18*. 84–91.

[2] Rohit Babbar and Bernhard Schölkopf. 2017. Dismec: Distributed sparse machines for extreme multi-label classification. In *WSDM'17*. 721–729.

[3] Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. In *EMNLP'19*. 3615–3620.

[4] Wei-Cheng Chang, Hsiang-Fu Yu, Kai Zhong, Yiming Yang, and Inderjit S Dhillon. 2020. Taming pretrained transformers for extreme multi-label text classification. In *KDD'20*. 3163–3171.

[5] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv:1412.3555* (2014).

[6] Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S Weld. 2020. SPECTER: Document-level Representation Learning using Citation-informed Transformers. In *ACL'20*. 2270–2282.

[7] Margaret H Coletti and Howard L Bleich. 2001. Medical subject headings used to search the biomedical literature. *JAMIA* 8, 4 (2001), 317–323.

[8] Kunal Dahiya, Deepak Saini, Anshul Mittal, Ankush Shaw, Kushal Dave, Akshay Soni, Himanshu Jain, Sumeet Agarwal, and Manik Varma. 2021. Deepxml: A deep extreme multi-label learning framework applied to short text documents. In *WSDM'21*. 31–39.

[9] Suyang Dai, Ronghui You, Zhiyong Lu, Xiaodi Huang, Hiroshi Mamitsuka, and Shanfeng Zhu. 2020. FullMeSH: improving large-scale MeSH indexing with full text. *Bioinformatics* 36, 5 (2020), 1533–1541.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In

NAACL-HLT'19. 4171–4186.
[11] Yuxiao Dong, Nitesh V Chawla, and Ananthram Swami. 2017. metapath2vec: Scalable representation learning for heterogeneous networks. In *KDD'17*. 135–144.
[12] Yuxiao Dong, Hao Ma, Zhihong Shen, and Kuansan Wang. 2017. A century of science: Globalization of scientific collaborations, citations, and innovations. In *KDD'17*. 1437–1446.
[13] Jorge E Hirsch. 2005. An index to quantify an individual's scientific research output. *PNAS* 102, 46 (2005), 16569–16572.
[14] Ziniu Hu, Yuxiao Dong, Kuansan Wang, and Yizhou Sun. 2020. Heterogeneous graph transformer. In *WWW'20*. 2704–2710.
[15] Himanshu Jain, Venkatesh Balasubramanian, Bhanu Chunduri, and Manik Varma. 2019. Slice: Scalable linear extreme classifiers trained on 100 million labels for related searches. In *WSDM'19*. 528–536.
[16] Himanshu Jain, Yashoteja Prabhu, and Manik Varma. 2016. Extreme multi-label loss functions for recommendation, tagging, ranking & other missing label applications. In *KDD'16*. 935–944.
[17] Ting Jiang, Deqing Wang, Leilei Sun, Huayi Yang, Zhengyang Zhao, and Fuzhen Zhuang. 2021. Lightxml: Transformer with dynamic negative sampling for high-performance extreme multi-label text classification. In *AAAI'21*. 7987–7994.
[18] Ching Jin, Yifang Ma, and Brian Uzzi. 2021. Scientific prizes and the extraordinary growth of scientific topics. *Nature Communications* 12, 1 (2021), 5619.
[19] Sujay Khandagale, Han Xiao, and Rohit Babbar. 2020. Bonsai: diverse and shallow trees for extreme multi-label classification. *Machine Learning* 109, 11 (2020), 2099–2119.
[20] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36, 4 (2020), 1234–1240.
[21] Jingzhou Liu, Wei-Cheng Chang, Yuexin Wu, and Yiming Yang. 2017. Deep learning for extreme multi-label text classification. In *SIGIR'17*. 115–124.
[22] Ke Liu, Shengwen Peng, Junqiu Wu, Chengxiang Zhai, Hiroshi Mamitsuka, and Shanfeng Zhu. 2015. MeSHLabeler: improving the accuracy of large-scale MeSH indexing by integrating diverse evidence. *Bioinformatics* 31, 12 (2015), i339–i347.
[23] Weiwei Liu, Haobo Wang, Xiaobo Shen, and Ivor W Tsang. 2021. The emerging trends of multi-label learning. *IEEE TPAMI* 44, 11 (2021), 7955–7974.
[24] Xiao Liu, Da Yin, Jingnan Zheng, Xingjian Zhang, Peng Zhang, Hongxia Yang, Yuxiao Dong, and Jie Tang. 2022. OAG-BERT: Towards a Unified Backbone Language Model for Academic Knowledge Services. In *KDD'22*. 3418–3428.
[25] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).
[26] Zhiyong Lu. 2011. PubMed and beyond: a survey of web tools for searching biomedical literature. *Database* 2011 (2011).
[27] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *JMLR* 9 (2008), 2579–2605.
[28] Dheeraj Mekala, Xinyang Zhang, and Jingbo Shang. 2020. META: Metadata-Empowered Weak Supervision for Text Classification. In *EMNLP'20*. 8351–8361.
[29] Anshul Mittal, Kunal Dahiya, Sheshansh Agrawal, Deepak Saini, Sumeet Agarwal, Purushottam Kar, and Manik Varma. 2021. Decaf: Deep extreme classification with label features. In *WSDM'21*. 49–57.
[30] Anshul Mittal, Noveen Sachdeva, Sheshansh Agrawal, Sumeet Agarwal, Purushottam Kar, and Manik Varma. 2021. ECLARE: Extreme Classification with Label Graph Correlations. In *WWW'21*. 3721–3732.
[31] Shengwen Peng, Ronghui You, Hongning Wang, Chengxiang Zhai, Hiroshi Mamitsuka, and Shanfeng Zhu. 2016. DeepMeSH: deep semantic representation for improving large-scale MeSH indexing. *Bioinformatics* 32, 12 (2016), i70–i79.
[32] Yashoteja Prabhu, Anil Kag, Shilpa Gopinath, Kunal Dahiya, Shrutendra Harsola, Rahul Agrawal, and Manik Varma. 2018. Extreme multi-label learning with label features for warm-start tagging, ranking & recommendation. In *WSDM'18*. 441–449.
[33] Yashoteja Prabhu, Anil Kag, Shrutendra Harsola, Rahul Agrawal, and Manik Varma. 2018. Parabel: Partitioned label trees for extreme classification with application to dynamic search advertising. In *WWW'18*. 993–1002.
[34] Yashoteja Prabhu and Manik Varma. 2014. Fastxml: A fast, accurate and stable tree-classifier for extreme multi-label learning. In *KDD'14*. 263–272.
[35] Martin Rosvall and Carl T Bergstrom. 2011. Multilevel compression of random walks on networks reveals hierarchical organization in large integrated systems. *PloS one* 6, 4 (2011), e18209.
[36] Deepak Saini, Arnav Kumar Jain, Kushal Dave, Jian Jiao, Amit Singh, Ruofei Zhang, and Manik Varma. 2021. GalaXC: Graph Neural Networks with Labelwise Attention for Extreme Classification. In *WWW'21*. 3733–3744.
[37] Zhihong Shen, Hao Ma, and Kuansan Wang. 2018. A Web-scale system for scientific knowledge exploration. In *ACL'18 System Demonstrations*. 87–92.
[38] Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June Hsu, and Kuansan Wang. 2015. An overview of microsoft academic service (mas) and applications. In *WWW'15 Companion*. 243–246.
[39] Yukihiro Tagami. 2017. Annexml: Approximate nearest neighbor search for extreme multi-label classification. In *KDD'17*. 455–464.
[40] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. 2008. Arnetminer: extraction and mining of academic social networks. In *KDD'08*. 990–998.
[41] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NIPS'17*. 5998–6008.
[42] Kuansan Wang, Zhihong Shen, Chiyuan Huang, Chieh-Han Wu, Yuxiao Dong, and Anshul Kanakia. 2020. Microsoft academic graph: When experts are not enough. *Quantitative Science Studies* 1, 1 (2020), 396–413.
[43] Boya Xie, Zhihong Shen, and Kuansan Wang. 2021. Is preprint the future of science? A thirty year journey of online preprint services. *arXiv preprint arXiv:2102.09066* (2021).
[44] Guangxu Xun, Kishlay Jha, Jianhui Sun, and Aidong Zhang. 2020. Correlation networks for extreme multi-label text classification. In *KDD'20*. 1074–1082.
[45] Guangxu Xun, Kishlay Jha, Ye Yuan, Yaqing Wang, and Aidong Zhang. 2019. MeSHProbeNet: a self-attentive probe net for MeSH indexing. *Bioinformatics* 35, 19 (2019), 3794–3802.
[46] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *NeurIPS'19*.
[47] Chenchen Ye, Linhai Zhang, Yulan He, Deyu Zhou, and Jie Wu. 2021. Beyond Text: Incorporating Metadata and Label Structure for Multi-Label Document Classification using Heterogeneous Graphs. In *EMNLP'21*. 3162–3171.
[48] Hui Ye, Zhiyu Chen, Da-Han Wang, and Brian Davison. 2020. Pretrained generalized autoregressive model with adaptive probabilistic label clusters for extreme multi-label text classification. In *ICML'20*. 10809–10819.
[49] Ian EH Yen, Xiangru Huang, Wei Dai, Pradeep Ravikumar, Inderjit Dhillon, and Eric Xing. 2017. Ppdsparse: A parallel primal-dual sparse method for extreme classification. In *KDD'17*. 545–553.
[50] Ian En-Hsu Yen, Xiangru Huang, Pradeep Ravikumar, Kai Zhong, and Inderjit Dhillon. 2016. Pd-sparse: A primal and dual sparse approach to extreme multiclass and multilabel classification. In *ICML'16*. 3069–3077.
[51] Yian Yin, Yuxiao Dong, Kuansan Wang, Dashun Wang, and Benjamin F Jones. 2022. Public use and public funding of science. *Nature Human Behaviour* 6, 10 (2022), 1344–1350.
[52] Ronghui You, Yuxuan Liu, Hiroshi Mamitsuka, and Shanfeng Zhu. 2021. BERTMeSH: deep contextual representation learning for large-scale high-performance MeSH indexing with full text. *Bioinformatics* 37, 5 (2021), 684–692.
[53] Ronghui You, Zihan Zhang, Ziye Wang, Suyang Dai, Hiroshi Mamitsuka, and Shanfeng Zhu. 2019. Attentionxml: Label tree-based attention-aware deep model for high-performance extreme multi-label text classification. *NeurIPS'19* (2019), 5820–5830.
[54] Hsiang-Fu Yu, Kai Zhong, Jiong Zhang, Wei-Cheng Chang, and Inderjit S Dhillon. 2022. PECOS: Prediction for enormous and correlated output spaces. *JMLR* 23, 98 (2022), 1–32.
[55] Fanjin Zhang, Xiao Liu, Jie Tang, Yuxiao Dong, Peiran Yao, Jie Zhang, Xiaotao Gu, Yan Wang, Bin Shao, Rui Li, et al. 2019. Oag: Toward linking large-scale heterogeneous entity graphs. In *KDD'19*. 2585–2595.
[56] Jiong Zhang, Wei-Cheng Chang, Hsiang-Fu Yu, and Inderjit Dhillon. 2021. Fast multi-resolution transformer fine-tuning for extreme multi-label text classification. In *NeurIPS'21*. 7267–7280.
[57] Yu Zhang, Xiusi Chen, Yu Meng, and Jiawei Han. 2021. Hierarchical Metadata-Aware Document Categorization under Weak Supervision. In *WSDM'21*. 770–778.
[58] Yu Zhang, Shweta Garg, Yu Meng, Xiusi Chen, and Jiawei Han. 2022. Motifclass: Weakly supervised text classification with higher-order metadata information. In *WSDM'22*. 1357–1367.
[59] 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.
[60] Yu Zhang, Zhihong Shen, Chieh-Han Wu, Boya Xie, Junheng Hao, Ye-Yi Wang, Kuansan Wang, and Jiawei Han. 2022. Metadata-Induced Contrastive Learning for Zero-Shot Multi-Label Text Classification. In *WWW'22*. 3162–3173.