META: Metadata-Empowered Weak Supervision for Text Classification

Dheeraj Mekala[1] **Xinyang Zhang**[2] **Jingbo Shang**[1,3]

[1] Department of Computer Science and Engineering, University of California San Diego, CA, USA

[2] Department of Computer Science, University of at Illinois Urbana-Champaign, IL, USA

[3] Halıcıoğlu Data Science Institute, University of California San Diego, CA, USA

[1,3] {dmekala, jshang}@ucsd.edu [2] xz43@illinois.edu

Abstract

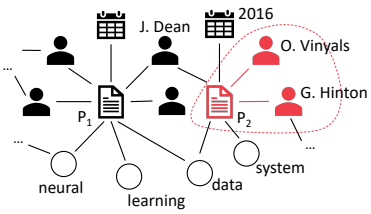
Recent advances in weakly supervised learning enable training high-quality text classifiers by only providing a few user-provided seed words. Existing methods mainly use text data alone to generate pseudo-labels despite the fact that metadata information (e.g., author and timestamp) is widely available across various domains. Strong label indicators exist in the metadata and it has been long overlooked mainly due to the following challenges: (1) metadata is multi-typed, requiring systematic modeling of different types and their combinations, (2) metadata is noisy, some metadata entities (e.g., authors, venues) are more compelling label indicators than others. In this paper, we propose a novel framework, META, which goes beyond the existing paradigm and leverages metadata as an additional source of weak supervision. Specifically, we organize the text data and metadata together into a text-rich network and adopt network motifs to capture appropriate combinations of metadata. Based on seed words, we rank and filter motif instances to distill highly label-indicative ones as "seed motifs", which provide additional weak supervision. Following a bootstrapping manner, we train the classifier and expand the seed words and seed motifs iteratively. Extensive experiments and case studies on real-world datasets demonstrate superior performance and significant advantages of leveraging metadata as weak supervision.

1 Introduction

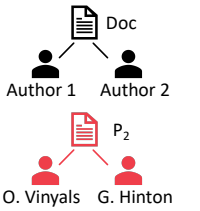
Weakly supervised text classification has recently gained much attention from the researchers because it reduces the burden of annotating the data. So far, the major source of weak supervision lies in text data itself (Agichtein and Gravano, 2000; Kuipers et al., 2006; Riloff et al., 2003; Tao et al., 2015; Meng et al., 2018; Mekala and Shang, 2020). These methods typically require a few user-provided seed words for each class as weak supervision. They expand seed words with generated pseudo labels and improve their text classifier in an iterative fashion.

Paper	Authors	Year	Category
P_1	G. Hinton, S. Osindero, YW. Teh	2006	ML
P_2	G. Hinton, O. Vinyals, J. Dean	2015	ML
P_3	J. Dean, S.Ghemawat	2008	Sys

(a) Examples of research papers with metadata.



(b) A text-rich network view of the papers.



(c) A motif pattern and a motif instance.

Figure 1: Text corpus, text-rich network, and motif.

Metadata information (e.g., author, published year) in addition to textual information, is widely available across various domains (e.g., news articles, social media posts, and scientific papers) and it could serve as a strong, complementary weak supervision source. Take a look at the research papers in Figure 1(a) as an example. It shall be learned in a data-driven manner that *G. Hinton* is a highly-reputed machine learning researcher, thus his presence is a strong indicator of a paper belonging to the *Machine Learning* category.

Distilling effective metadata for weak supervision faces several major challenges. Metadata is often multi-typed, each type and the type combinations could have very different semantics and may not be equally important. Moreover, even entities within a single metadata type could be noisy. Continuing our example in Figure 1(a), we shall notice that *year* is less helpful than an *author* to do classification. Among the authors, *J. Dean* might be an important figure but has research interests

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 8351–8361,
November 16–20, 2020.

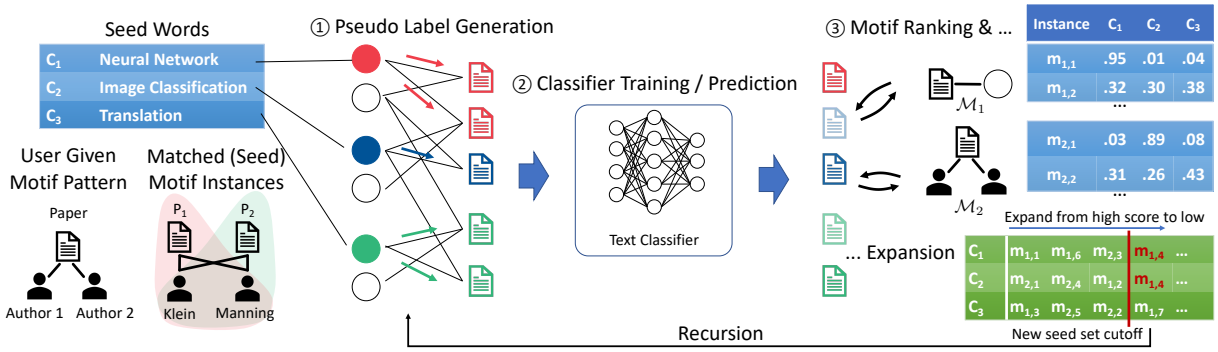


Figure 2: Our META framework. In each iteration, we generate pseudo labels for documents, train the text classifier, and rank all words and motif instances in a unified ranking framework. We then expand seed sets until an automatic cutoff is reached. The quality of the classifier and the seed sets are improved through iterations.

spanning across different domains. However, if we join the *author* with *year*, it carries more accurate semantics, and we may discover *J. Dean* has more interest in machine learning in recent years, thus becoming highly label-indicative.

Bearing the challenges in mind, we propose META, a principled framework for metadata-empowered weakly-supervised text classification. As illustrated in Figure 1 and Figure 2, we first organize the text data and metadata together into a text-rich network. The network structure gives us a holistic view of the corpus and enables us to rank and select useful metadata entities. We leverage motif patterns (Benson et al., 2016; Milo et al., 2002; Shang et al., 2020) to model typed metadata as well as their combinations. A motif pattern is a subgraph pattern at the meta-level that captures higher-order connections and the semantics represented by these connections. It serves as a useful tool to model typed edges, typed paths (a.k.a. meta-paths) (Sun et al., 2011), and higher-order structures in the network. With little effort, users can specify a few possibly useful motif patterns as input to our model. We develop a unified, principled ranking mechanism to select label-indicative motif instances and words, forming expanded weak supervision. Note that, such instance-level selection process also implicitly refines the motif patterns, ensuring the robust performance of META even when irrelevant motif patterns exist in input. It is worth a mention that META is compatible with any text classifiers.

Our contributions are summarized as follows:

- We explore to incorporate metadata information as an additional source of weak supervision for text classification along with seed words.
- We propose a novel framework META, which introduces motif patterns to capture the high-order combinations among different types of metadata and conducts a unified ranking and selection of label-indicative motif instances and words.
- We conduct experiments on two real-world datasets. The results and case studies demonstrate the superiority of incorporating metadata as parts of weak supervision and verify the effectiveness of META.

Reproducibility. Our code is made publicly available at GitHub[1].

2 Preliminaries

2.1 Documents as Text-rich Network

Given a collection of n text documents $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_n\}$, and their corresponding metadata, we propose to organize them into a *text-rich network*, as illustrated in Figure 1(b). A text-rich network is a heterogeneous network with documents, words, different types of metadata as nodes, and their associations as edges. For example, our text-rich network for research papers has papers, words, authors, and publication years as nodes. Each paper is connected to its associated words and metadata nodes. Such a network provides a holistic and structured representation of the input.

2.2 Seed Words and Motif Patterns

Users are asked to provide a few seed words $\mathcal{S} = \{\mathcal{S}_1^w, \mathcal{S}_2^w, \ldots, \mathcal{S}_l^w\}$ for each of l classes (i.e., $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_l$) in our classification problem, as well as k motif patterns $\{\mathcal{M}_1, \mathcal{M}_2, \ldots, \mathcal{M}_k\}$. Motif patterns are sub-graph patterns at the meta-level (i.e., every node is abstracted by its type). They are able to capture semantics and higher-order interconnections among nodes. A motif instance is a

[1] https://github.com/dheeraj7596/META

sub-graph instance in the graph that follows a motif pattern. Figure 1 presents an example of a motif pattern that captures co-authorship and a motif instance following this motif pattern. In this paper, we discover seed motif instances for each class label, denoted as $\{\mathcal{S}_1^m, \mathcal{S}_2^m, \ldots, \mathcal{S}_l^m\}$.

2.3 Problem Formulation

Given the text-rich network and user-provided seed words and motif patterns as input, we aim to build a high-quality document classifier, assigning one class label $\mathcal{C}_j$ to each document $\mathcal{D}_i$.

3 Our META Framework

As shown in Figure 2, META is an iterative framework, generating pseudo labels and training the text classifier alternatively, similar to many other weakly supervised text classification methods (Kuipers et al., 2006; Tao et al., 2015; Meng et al., 2018). One iteration in META consists of the following steps:

- Generate pseudo labels based on the seeds;
- Train a text classifier based on pseudo labels;
- Rank and select words and motif instances to expand the seeds.

We repeat these steps iteratively. We denote the number of iterations as T, which is the only hyper-parameter in our framework.

The novelty of META mainly lies in integrating two sources of weak supervisions, seed motif instances, and seed words. Given each motif instance m or each word w, for each label l, we estimate a *ranking score* $\mathcal{R}_{m,l}$ or $\mathcal{R}_{w,l}$ ranging between 0 and 1, measuring how label-indicative it is to the particular label l. Such ranking scores are utilized to select new seed motif instances and seed words. Note that, while this selection is conducted at the instance level, it also selects motif patterns implicitly and therefore ensures robust performance when users provide some irrelevant motif patterns.

3.1 Pseudo Labels and Text Classifier

Based on seed words, seed motif instances, and their respective ranking scores for each class, we generate pseudo labels for unlabeled text documents and train a classifier based on these pseudo labels. In the first iteration, we have no seed motif instances and the ranking score is 1 for all seed words.

Pseudo-Label Generation. Suppose we have seed word sets $\mathcal{S}_{1..l}^w$ and seed motif instance sets $\mathcal{S}_{1..l}^m$ for all l labels, we generate pseudo labels using a simple yet effective count-based technique. Specifically, given a document $\mathcal{D}_i$, the probability that it belongs to the class l is proportional to the aggregated ranking scores of its respective seed words and seed motif instances.

$$P(l|\mathcal{D}_i) \propto \sum_{w \in \mathcal{D}_i \cap \mathcal{S}_l^w} f_{\mathcal{D}_i,w} \cdot \mathcal{R}_{w,l} + \sum_{m \in \mathcal{D}_i \cap \mathcal{S}_l^m} \mathcal{R}_{m,l}$$

where $f_{\mathcal{D}_i,w}$ is the term frequency of the word w in document $\mathcal{D}_i$. The pseudo label of document $\mathcal{D}_i$ is then assigned as follows:

$$l(\mathcal{D}_i) = \arg\max_l P(l|\mathcal{D}_i)$$

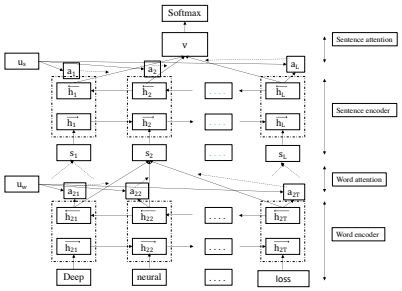


Figure 3: HAN Classifier used in our META.

Document Classifier. Our framework is compatible with any text classification model as a classifier. We use Hierarchical Attention Networks (HAN) (Yang et al., 2016) as the classifier. HAN is designed to capture the hierarchical document structure i.e. words – sentences – documents. As illustrated in Figure 3, HAN performs attention first on the sentences in the document to find the important sentence in a document and on the words in the sentence to identify important words in a sentence. We train a HAN model on unlabeled documents with the generated pseudo-labels. For the document $\mathcal{D}_i$, it estimates the probability $\hat{Y}_{i,l}$ for each class l. Such predicted distributions are used in the expansion of seed words and motifs.

3.2 Unified Seed Ranking and Expansion

Once the text classifier is trained, we rank words and motif instances together for each class. Then, we expand the seed sets by adding top-ranked words and motif instances. This improves the quality of the weak supervision over iterations, thereby improving the text classifier. We present our design of the unified ranking and expansion as follows.

Ranking Score Design. An ideal seed word or motif instance for a particular class should be highly relevant and highly exclusive to this class. So an

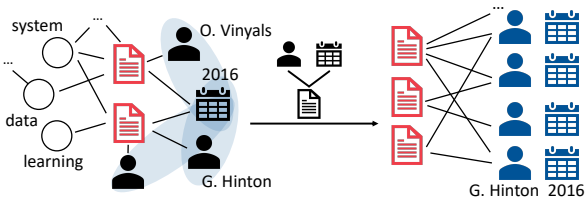


Figure 4: Using motif patterns, we construct bipartite graphs from the text-rich network linking documents to their respective motif instances.

effective ranking score must quantify relevance and exclusiveness. Such a ranking score for words alone has been explored by previous studies (Tao et al., 2015; Mekala and Shang, 2020), typically based on similarity and frequency-based metrics. In this paper, we have motif instances in addition to words, therefore, we build upon the text-rich network to unify the ranking process.

Given k user-provided motif patterns $\mathcal{M}_1, \ldots, \mathcal{M}_k$ and the text-rich network $\mathcal{G}$, we construct k bipartite graphs $\mathcal{G}_1^B, \ldots, \mathcal{G}_k^B$, one for each motif pattern (see Figure 4). In the i-th bipartite graph $\mathcal{G}_i^B$, the node set contains two parts: (1) all documents and (2) all motif instances following the motif pattern $\mathcal{M}_i$ in the text-rich network $\mathcal{G}$; The edges in the graph $\mathcal{G}_i^B$ connect the documents to the motif instances which are subsets of the metadata associated with the documents.

For the sake of simplicity, we introduce one more motif pattern, document–word. It makes words a special case of motif instances, and one can easily construct a similar bipartite graph for words. Therefore, in the rest of this section, we use motif instances to explain our ranking score design.

For each motif pattern $\mathcal{M}$, we conduct one personalized random walk on its corresponding bipartite graph $\mathcal{G}^B$ for each label l. Specifically, we normalize each column of the adjacency matrix of the bipartite graph $\mathcal{G}^B$ by the degree of its respective node, resulting in the transition matrix $\mathbf{W}$. Suppose $\mathbf{p}_{l,u}$ represents the personalized PageRank (PPR) score of each node u for each label l, we initialize the PPR score of each document node to $\hat{Y}_{i,l}$ and PPR score of each motif instance node to 0. This initialization ensures that a random walk starts from a document node and since $\mathcal{G}^B$ is bipartite, it ends at a motif instance node. We iteratively update the PPR scores as follows:

$$\mathbf{p}_l^{(t+1)} \leftarrow \mathbf{W}\mathbf{p}_l^{(t)}$$

Since each document node is initialized with probabilities corresponding to l and the random walk starts from a document node and ends at a motif instance node, this can be viewed as a label propagation problem. Based on the previous work in label propagation (Hensley et al., 2015), similar nodes are more likely to form edges and the PPR score is used to measure the similarity. Therefore, we believe that $\mathbf{p}_{l,m}$ reflects the *relevance* of a motif instance m to the particular class label l.

Though the absolute values of PPR scores are quite small, their relative magnitude conveys their affinity towards a label. Therefore, we normalize these PPR scores into a distribution, resulting in the ranking scores. Mathematically, for a label l, the ranking score of a motif instance m is:

$$\mathcal{R}_{m,l} = \frac{\mathbf{p}_{l,m}}{\sum_{l' \in \mathcal{C}} \mathbf{p}_{l',m}}$$

If a motif instance has similar relevance to multiple labels, the ranking score distribution becomes flat irrespective of the magnitude of its respective PPR scores. From this, we realize that our ranking score also quantifies *exclusiveness*, which is an essential characteristic of a highly label-indicative term.

Based on this ranking score, we rank words and motif instances in a unified manner and expand the seed word set and seed motifs set.

Expansion. Given the ranking scores of all words and motif instances for every label, we expand the seed words and seed motifs simultaneously for all labels. Intuitively, a highly label-indicative motif instance would not belong to the seed sets of multiple labels. Therefore, when any motif instance is expanded to seed sets of multiple classes, we stop the expansion of motif instances of the corresponding motif pattern. Also, we set a hard threshold of $\frac{1}{|\mathcal{C}|}$, where $|\mathcal{C}|$ is the number of classes, on ranking scores for those added motif instances. In this way, the number of new seed words and seed motif instances is decided by the method automatically. It is worth mentioning that our expansion here is adaptive and every label may have a different number of seeds. Note that, in the first iteration, pseudo labels are generated using only seed words but ranking scores are obtained for all words and motif instances. The highly ranked motif instances and words are used as seeds in further iterations.

After expanding the seed sets for every label, we generate pseudo labels and train the classifier. This process is repeated iteratively for T iterations.

Table 1: Dataset statistics.

Dataset	# Docs	# Classes	Avg Doc Len
DBLP	38,128	9	893
Book Graph	33,594	8	620

4 Experiments

In this section, we evaluate META and compare it with existing techniques on two real-world datasets in a weakly supervised classification setting.

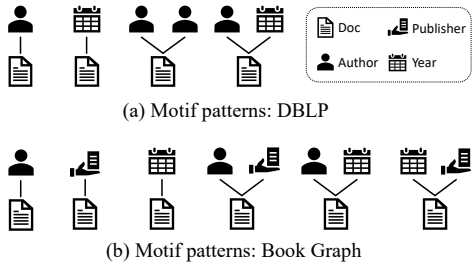
4.1 Experimental Settings

Datasets. We conduct experiments on the DBLP dataset (Tang et al., 2008) and the Book Graph dataset (Wan and McAuley, 2018; Wan et al., 2019). The dataset statistics are shown in Table 1. The details of the datasets are mentioned below.

- **DBLP dataset:** The DBLP dataset contains a comprehensive set of research papers in computer science. We select 38,128 papers published in flagship venues. In addition to text data, it has information about authors, published year, and venue for each paper. There are 9,300 distinct authors and 42 distinct years. For each paper, we annotate its research area largely based on its venue as the classification objective[2]. Therefore, in our experiments, we drop the venue information to ensure a fair comparison.
- **Book Graph dataset:** The Book Graph dataset is a collection of the description of books, user-book interactions, and users' book reviews collected from a popular online book review website named Goodreads[3]. We select books belonging to eight popular genres[4]. The genre of a book is viewed as the label to be predicted. The total number of books selected is 33,594. We use the title and description of a book as text data and author, publisher, and year as metadata. In total, there are 22,145 distinct authors, 5,186 distinct publishers, and 136 distinct years.

Motif Patterns. The motif patterns we used as metadata information for DBLP and Book Graph datasets are shown in Figure 5.

Seed Words. The seed words are obtained as follows: we asked 5 human experts to recommend 5 seed words for each class and selected the final seed words based on majority voting i.e. (> 3 recommendations).



(a) Motif patterns: DBLP

(b) Motif patterns: Book Graph

Figure 5: Motif Patterns used in Experiments.

Evaluation Metrics. Both datasets are imbalanced with respect to the label distribution. Being aware of this fact, we adopt micro- and macro-F_1 scores as evaluation metrics.

Implementation Details. To make the model robust to multi-word phrases as supervision, we extract phrases using Autophrase (Liu et al., 2015; Shang et al., 2018). We set the word vector dimension to be 100 for all the methods that use word embeddings. We set the number of iterations parameter for META to 9.

4.2 Compared Methods

We compare our proposed method with a wide range of methods described below:

- **IR-TF-IDF** treats seed words as a query. It computes the relevance of a document to a class by aggregating the TF-IDF values of its seed words. Each document is assigned the label which is the most relevant to this document.
- **Word2Vec** learns word vector representations (Mikolov et al., 2013) for all words in the corpus. It computes label representations by aggregating the word vectors of all its seed words. Each document is assigned the label whose cosine similarity with this document is maximum.
- **Doc2Cube** (Tao et al., 2015) considers label surface names as seed set and performs multi-dimensional document classification by learning dimension-aware embedding.
- **WeSTClass** (Meng et al., 2018) leverages seed words to generate bag-of-words pseudo documents for neural model pre-training and then bootstraps the model on unlabeled data. Specifically, we compare with WeSTClass-CNN which is the best configuration under our setting. We use the public implementations of WeSTClass[5]

[2]Classes in DBLP: (1) computer vision, (2) computational linguistics, (3) biomedical engineering, (4) software engineering, (5) graphics, (6) data mining, (7) security and cryptography, (8) signal processing, (9) robotics, and (10) theory.

[3]https://www.goodreads.com/

[4]Classes in Book Graph: (1) children, (2) graphic comics, (3) paranormal fantasy, (4) history & biography, (5) crime, mystery thriller, (6) poetry, (7) romance, and (8) young adult.

[5]https://github.com/yumeng5/WeSTClass

Table 2: Evaluation Results on Two Datasets. ++ represents that the input is metadata-augmented.

Methods	DBLP		Books Graph	
	Mi-F_1	Ma-F_1	Mi-F_1	Ma-F_1
IR-TF-IDF	0.19	0.20	0.24	0.29
Word2Vec	0.23	0.22	0.28	0.26
Doc2Cube	0.37	0.36	0.33	0.31
WeSTClass	0.58	0.53	0.42	0.41
Metapath2Vec	0.64	0.61	0.47	0.48
IR-TF-IDF++	0.19	0.20	0.24	0.29
Word2Vec++	0.24	0.21	0.26	0.25
Doc2Cube++	0.40	0.38	0.36	0.33
WeSTClass++	0.60	0.55	0.47	0.43
META	**0.66**	**0.63**	0.62	**0.63**
META-CNN	0.61	0.58	0.54	0.55
META-BERT	0.64	0.61	**0.63**	**0.63**
META-NoMeta	0.61	0.58	0.58	0.58
META-CNN-NoMeta	0.56	0.53	0.53	0.53
META-BERT-NoMeta	0.58	0.57	0.60	0.60
HAN-Sup	0.75	0.72	0.77	0.76
HAN-Sup++	0.79	0.77	0.81	0.81

with the hyperparameters mentioned in the paper.

- **Metapath2Vec** (Dong et al., 2017) learns node representations in the text-rich network using meta-path-guided random walks by capturing the structural and semantic correlations of differently typed nodes. We use the first two motif patterns in Figure 5(a) and the first three motif patterns in Figure 5(b) as meta-paths because the rest cannot be represented as meta-paths. We generate pseudo-labels using the seed words and train a logistic regression classifier with document nodes representations as input to predict the labels.

We denote our framework with HAN classifier as **META**, with CNN classifier as **META-CNN**, and with BERT(bert-base-uncased) classifier as **META-BERT**. We also compare with their respective ablated versions **META-NoMeta**, **META-CNN-NoMeta**, **META-BERT-NoMeta** where metadata information is not expanded and not considered while generating pseudo labels.

For a fair comparison, we also present results of all the baselines on the **metadata-augmented** datasets, where a token for every relevant motif instance is appended to the text data of a document. This is denoted by **++** in Table 2, e.g., WeSTClass++ represents the performance of WeSTClass on metadata-augmented datasets.

We also present the performance of HAN in a supervised setting which is denoted as **HAN-Sup**. The results of HAN-Sup reported are on the test set which follows an 80-10-10 train-dev-test split.

4.3 Performance Comparison

The evaluation results of all methods are summarized in Table 2. We can observe that our proposed framework outperforms all the compared weakly supervised methods. We discuss the effectiveness of our proposed META as follows:

- META achieves the best performance among all the compared weakly supervised methods with significant margins. By extracting the highly label-indicative motif instances along with words and using them together in pseudo label generation, META successfully leverages metadata information and achieves superior performance.
- We observe that the performance of META is better than all the compared weakly supervised models on metadata-augmented datasets. By comparing those **++** methods with their text-only counterparts, one can easily observe that adding metadata in text classification is indeed helpful. However, META does not restrict to single metadata types and goes beyond by employing motif patterns to capture the metadata information. It is successful in identifying the appropriate label-indicative metadata combinations and therefore achieves even better performance.
- The comparison between META and Metapath2Vec demonstrates the advantages of motif patterns over the meta-paths. For example, on the Book Graph dataset, the last three motif patterns in Figure 5(b) cannot be represented through meta-paths and this significantly affects the performance. It's also worth mentioning that Metapath2Vec cannot handle new documents directly without re-training the embedding whereas our framework can directly predict without any additional effort.
- The comparison between META and the ablation method META-NoMeta demonstrates the effectiveness of our motif instance expansion. For example, on the Book Graph dataset, the motif instance expansion improves the micro-F1 score from 0.58 to 0.62 and macro-F1 score from 0.58 to 0.63, which are quite significant.
- The comparison between META-CNN, META-BERT, and their respective ablated versions META-CNN-NoMeta, META-BERT-NoMeta demonstrate that our proposed approach provides significant additive gains to different classifiers and thereby showing the effectiveness of leveraging metadata information as an additional source of weak supervision.

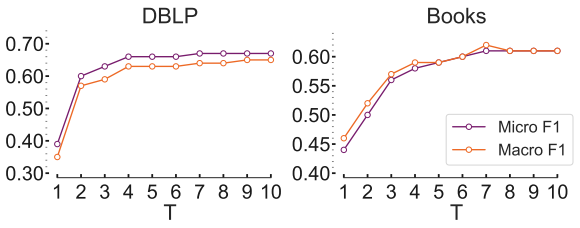


Figure 6: Micro- and Macro-F_1 scores w.r.t. the number of iterations.

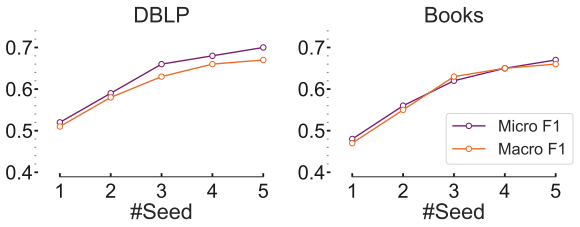


Figure 7: Micro- and Macro-F_1 scores w.r.t. the number of seed words.

- The comparison between META and HAN-Sup demonstrates that META is effective in decreasing the gap between the performance of the weakly supervised and supervised settings.

4.4 Parameter Study

The only hyper-parameter in our framework META is T, the number of iterations. We experiment on both datasets to study the effect of the number of iterations on the performance. The plots of micro-F1 score and macro-F1 score with respect to the number of iterations are shown in Figure 6. We observe that the performance increases initially and gets gradually converged by 6 or 7 iterations. We also observe that the expanded seed words and seed motifs have become almost unchanged. While there is some fluctuation, a reasonably large T, such as $T = 9$ or $T = 10$, is recommended.

4.5 Number of Seed Words

We vary the number of seed words per class and plot the performance in Figure 7. We observe that the performance increases as the number of seed words increase, which is generally intuitive. For reasonable performance, we observe that three seed words are sufficient.

4.6 Case Study

We present case studies to showcase the effectiveness of our framework in addressing the challenges of leveraging metadata.

Leveraging Metadata Combinations. Table 3 shows a few samples of expanded motif instances.

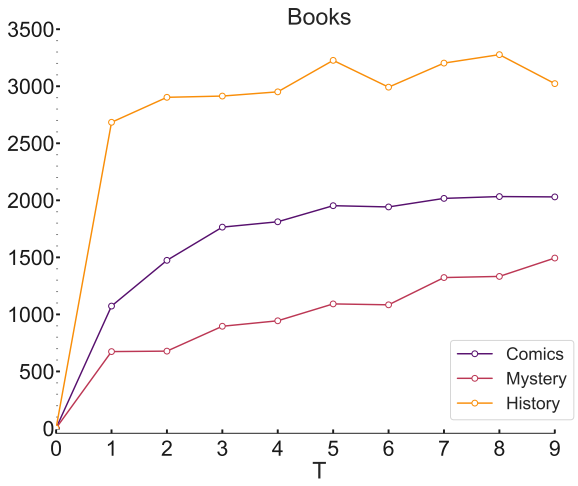


Figure 8: Number of seed words w.r.t. the number of iterations

First, let's take a look at motif instances related to authors and publishers. We can observe that strong label-indicative authors and publishers are mined accurately. For example, *Marvel*, a widely known comics publisher, is present in the expanded publishers for *comics* genre; A classic American poet *E. Dickinson* is successfully identified as label-indicative for *poetry* genre.

Note that, the author *N. Gaiman* (in blue) who has written books in multiple genres including comic books, graphic novels, etc., is not a label-indicative author for any of these categories, because he is not exclusive to any one category, which is accurately captured by our framework. However, his works in various genres together with their respective publisher information form a unique label-indicative pattern which is reflected by the "Author-Publisher" motif pattern.

Now, adding *year* metadata into the loop, although "Year-Document" is a user-provided motif pattern, META identifies that *year* information alone is not much helpful in classification. This demonstrates the robustness of our framework when users provide some irrelevant motif patterns. However, if we combine author information with year, it then carries more accurate semantics, and we may discover that *N.Gaiman* had authored more children's books in early 2000, thus becoming highly label-indicative.

Eliminating Noise in Metadata. Table 4 presents the percentage of motif instances expanded out of the total motif instances following a motif pattern, for every label. One can observe that META actually prunes out many motif instances, as the final selection ratio is far less than 100%.

Table 3: Case Study: Expanded motif instances.

Expanded motif instances of Book Graph dataset					
Class	**Author**	**Publisher**	**Author-Publisher**	**Year**	**Author-Year**
children	Z. Fraillon, K. Argent	Brighter Child, HarperCollins Children's Books	(**N. Gaiman**, Bloomsbury UK) (M. Fox, Penguin Australia)	N/A	(**N. Gaiman**, 2004) (S. Blackall, 2010)
comics	F. Teran, B. Kane	Marvel, Titan Books Ltd	(**N. Gaiman**, Marvel) (T. McFarlane, Marvel Comics)	N/A	(T. Hairsine, 2013) (A. Sinclair, 2009)
fantasy	J. Barne, S. Dubbin	DAW Books, Inc., Edge Publishing	(W. King, Titan Books Ltd) (G.J. Grant, Prime Books)	N/A	(G.J. Grant, 2012) (M. Lingen, 2012)
poetry	B. Guest, E. Dickinson	Shearsman Books, Souvenir Press	(**N. Gaiman**, MagicPress) (R. Browning, Wordsworth Editions)	1692, 1914	(E. Dickinson, 1959) (J. McCrae, 1929)

Table 4: Case Study: Percentage of motif instances expanded for Book Graph dataset. **A** stands for author, **P** for publisher and **Y** for year.

Percentage of motif instances expanded						
Label	**A**	**P**	**Y**	**A-P**	**P-Y**	**A-Y**
children	5.12	9.42	0	9.21	12.73	9.68
comics	4.91	1.33	0	9.52	1.48	14.11
fantasy	6.2	2.8	0	13.1	2.95	10.97
history	4.31	10.5	6.12	8.1	11.8	7.94
mystery	4.11	8.6	3.67	9.8	11.04	9.59
poetry	6.8	9.2	15.4	10	8.17	9.11
romance	5.6	13.5	1.47	9.6	12.28	9.19
y. adult	3.52	13.7	2.2	9.1	15.04	9.32

Table 5: Expanded seed words of *comics*, *history*, and *mystery* classes in Books dataset.

Expanded seed words	
Label	**Seed words**
comics	batman, superman, marvel, mary-jane, general zod
history	history, world war, world war ii, political science
mystery	serial killer, sherlock holmes, inspector lestrade

For the "**Y**ear-Document" motif pattern, we observe that its motif instances are only expanded for a few genres, which is generally intuitive. For example, one can see that a significant percentage of "Year-Document" motif instances expanded for *history* and *poetry*. After a closer inspection, we find that the expanded years were concentrated between the late 1800 and early 1900, thus developing an affinity for this time period.

One can also observe that the percentage of motif instances following the "**P**ublisher-Document" motif pattern expanded varies for different labels, ranging from 1 to 13.5. This illustrates that our expansion is adaptive.

Seed words Expansion. Figure 7 shows the number of seed words expanded after each iteration for *comics*, *hystory*, and *mystery* classes in Books dataset. We observe that the number varies for each label because of our data-driven, adaptive thresholds, which is different for each label.

One can also observe that the the number increases over iterations and gets almost stagnated at the end, indicating that the seed sets are getting refined and converged. A few examples of expanded seed words are shown in Table 5.

5 Related Work

We review the literature about (1) weakly supervised text classification methods, (2) text classification with metadata, and (3) document classifiers.

5.1 Weakly Supervised Text Classification

Due to the training data bottleneck in supervised classification, weakly supervised classification has recently attracted much attention from researchers. The majority of weakly supervised classification techniques require seeds in various forms, including label surface names (Li et al., 2018; Song and Roth, 2014; Tao et al., 2015), label-indicative words (Chang et al., 2008; Meng et al., 2018; Tao et al., 2015; Mekala and Shang, 2020), and labeled-documents (Tang et al., 2015b; Xu et al., 2017; Miyato et al., 2016; Meng et al., 2018).

Dataless (Song and Roth, 2014) considers label surface names as seeds and classifies documents by embedding both labels and documents in a semantic space and computing semantic similarity between a document and a potential label; Along similar lines, Doc2Cube (Tao et al., 2015) expands label-indicative words using label surface names and performs multi-dimensional document classification by learning dimension-aware embedding; WeSTClass (Meng et al., 2018) considers both word-level and document level supervision sources. It first generates bag-of-words pseudo documents for neural model pre-training, then bootstraps the model on unlabeled data. This method is later extended to a hierarchical setting with a pre-defined hierarchy (Meng et al., 2019); ConWea (Mekala and Shang, 2020) leverages contextualized representation techniques to provide contextualized

weak supervision for text classification.

However, all these techniques consider only the text data and don't leverage metadata information for classification. In this paper, we focus on user-provided seed words and mine label-indicative words and metadata in an iterative manner.

5.2 Text Classification with Metadata

Previous studies try to incorporate metadata information to improve the performance of the classifier. Tang et al. (2015a) and Chen et al. (2016) consider the user and product information as metadata for document-level sentiment classification; Rosen-Zvi et al. (2012) use author information for paper classification; Zhang et al. (2017) employ user biography data for tweet localization. However, all these frameworks are in a supervised setting and use fixed metadata types for each task whereas our method is generalized for different metadata types and multiple metadata combinations.

Another way to leverage metadata for text understanding is to organize the corpus into a heterogeneous information network. A straightforward approach is to obtain document representations using their respective meta-path guided node embeddings (Dong et al., 2017; Shang et al., 2016) and train a classifier. However, higher-order connectivity cannot be captured by meta-paths and this approach can't handle new documents directly without re-training the embeddings. Recently, Zhang et al. (2020) proposed a minimally supervised framework to categorize text with metadata. However, they require labeled documents as supervision and they only consider typed edges in the model. Network motifs (Milo et al., 2002) can capture higher-order connectivity and have been proved fundamental in complex real-word networks across various domains (Benson et al., 2016). Shang et al. (2020) leveraged motifs for topic taxonomy construction in an unsupervised setting. Our proposed method mines highly label-indicative metadata information with a unified motif and word ranking framework, and effectively expands weak supervision to improve document classification.

5.3 Document classifier

Document classification has been a long-studied problem in Natural Language Processing. CNN-based classifiers (Kim, 2014; Johnson and Zhang, 2014; Lai et al., 2015), RNN-based classifiers (Socher et al., 2013) achieve competitive performance. Yang et al. (2016) proposed Hierarchical Attention Network (HAN) for document classification that performs attention first on the sentences in the document, and on the words in the sentence to find the most important sentences and words in a document. Though our framework uses HAN as the document classifier, it is also compatible with all the above-mentioned text classifiers. We choose HAN for the demonstration purpose.

6 Conclusion and Future Work

In this paper, we propose META, a novel framework that leverages metadata information as an additional source of weak supervision and incorporates it into the classification framework. Our method organizes the text data and metadata together into a text-rich network and employs motif patterns to capture appropriate metadata combinations. Using the initial user-provided seed words and motif patterns, our method generates pseudo labels, trains classifier, and ranks and filters highly label-indicative words, motifs in a unified manner and adds them to their respective seed set. Experimental results and case studies demonstrate that our model outperforms previous methods significantly, thereby signifying the advantages of leveraging metadata as weak supervision.

In the future, we are interested in effectively integrating different forms of supervision including annotated documents. Also, we only consider positively label-indicative metadata combinations currently. There should be negatively label-indicative combinations as well which can eliminate some classes from potential labels. This is another potential direction for the extension of our method.

Acknowledgements

We thank anonymous reviewers and program chairs for their valuable and insightful feedback. The research was sponsored in part by National Science Foundation CA-2040727. Any opinions, findings, and conclusions or recommendations expressed herein are those of the authors and should not be interpreted as necessarily representing the views, either expressed or implied, of the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for government purposes notwithstanding any copyright annotation hereon.

References

Eugene Agichtein and Luis Gravano. 2000. Snowball: Extracting relations from large plain-text collections. In *Proceedings of the fifth ACM conference on Digital libraries*, pages 85–94.

Austin R Benson, David F Gleich, and Jure Leskovec. 2016. Higher-order organization of complex networks. *Science*, 353(6295):163–166.

Ming-Wei Chang, Lev-Arie Ratinov, Dan Roth, and Vivek Srikumar. 2008. Importance of semantic representation: Dataless classification. In *Aaai*, volume 2, pages 830–835.

Huimin Chen, Maosong Sun, Cunchao Tu, Yankai Lin, and Zhiyuan Liu. 2016. Neural sentiment classification with user and product attention. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1650–1659.

Yuxiao Dong, Nitesh V Chawla, and Ananthram Swami. 2017. metapath2vec: Scalable representation learning for heterogeneous networks. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 135–144.

A. Hensley, A. Doboli, R. Mangoubi, and S. Doboli. 2015. Generalized label propagation. In *2015 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Rie Johnson and Tong Zhang. 2014. Effective use of word order for text categorization with convolutional neural networks. *arXiv preprint arXiv:1412.1058*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

Benjamin J Kuipers, Patrick Beeson, Joseph Modayil, and Jefferson Provost. 2006. Bootstrap learning of foundational representations. *Connection Science*, 18(2):145–158.

Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent convolutional neural networks for text classification. In *Twenty-ninth AAAI conference on artificial intelligence*.

Keqian Li, Hanwen Zha, Yu Su, and Xifeng Yan. 2018. Unsupervised neural categorization for scientific publications. In *SIAM Data Mining*, pages 37–45. SIAM.

Jialu Liu, Jingbo Shang, Chi Wang, Xiang Ren, and Jiawei Han. 2015. Mining quality phrases from massive text corpora. In *Proceedings of the 2015 ACM SIGMOD International Conference on Management of Data*, pages 1729–1744.

Dheeraj Mekala and Jingbo Shang. 2020. Contextualized weak supervision for text classification. *arXiv preprint arXiv:1612.06778*.

Yu Meng, Jiaming Shen, Chao Zhang, and Jiawei Han. 2018. Weakly-supervised neural text classification. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 983–992. ACM.

Yu Meng, Jiaming Shen, Chao Zhang, and Jiawei Han. 2019. Weakly-supervised hierarchical text classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6826–6833.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. 2002. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827.

Takeru Miyato, Andrew M Dai, and Ian Goodfellow. 2016. Adversarial training methods for semi-supervised text classification. *arXiv:1605.07725*.

Ellen Riloff, Janyce Wiebe, and Theresa Wilson. 2003. Learning subjective nouns using extraction pattern bootstrapping. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 25–32. Association for Computational Linguistics.

Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. 2012. The author-topic model for authors and documents. *arXiv preprint arXiv:1207.4169*.

Jingbo Shang, Jialu Liu, Meng Jiang, Xiang Ren, Clare R Voss, and Jiawei Han. 2018. Automated phrase mining from massive text corpora. *IEEE Transactions on Knowledge and Data Engineering*, 30(10):1825–1837.

Jingbo Shang, Meng Qu, Jialu Liu, Lance M Kaplan, Jiawei Han, and Jian Peng. 2016. Meta-path guided embedding for similarity search in large-scale heterogeneous information networks. *arXiv preprint arXiv:1610.09769*.

Jingbo Shang, Xinyang Zhang, Liyuan Liu, Sha Li, and Jiawei Han. 2020. Nettaxo: Automated topic taxonomy construction from text-rich network. In *Proceedings of The Web Conference 2020*, pages 1908–1919.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Yangqiu Song and Dan Roth. 2014. On dataless hierarchical text classification. In *AAAI*.

Yizhou Sun, Jiawei Han, Xifeng Yan, Philip S Yu, and Tianyi Wu. 2011. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. *Proceedings of the VLDB Endowment*, 4(11):992–1003.

Duyu Tang, Bing Qin, and Ting Liu. 2015a. Learning semantic representations of users and products for document level sentiment classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1014–1023.

Jian Tang, Meng Qu, and Qiaozhu Mei. 2015b. Pte: Predictive text embedding through large-scale heterogeneous text networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1165–1174. ACM.

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*, pages 990–998.

Fangbo Tao, Chao Zhang, Xiusi Chen, Meng Jiang, Tim Hanratty, Lance Kaplan, and Jiawei Han. 2015. Doc2cube: Automated document allocation to text cube via dimension-aware joint embedding. *Dimension*, 2016:2017.

Mengting Wan and Julian J. McAuley. 2018. Item recommendation on monotonic behavior chains. In *Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018, Vancouver, BC, Canada, October 2-7, 2018*, pages 86–94. ACM.

Mengting Wan, Rishabh Misra, Ndapa Nakashole, and Julian J. McAuley. 2019. Fine-grained spoiler detection from large-scale review corpora. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2605–2610. Association for Computational Linguistics.

Weidi Xu, Haoze Sun, Chao Deng, and Ying Tan. 2017. Variational autoencoder for semi-supervised text classification. In *AAAI*.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489.

Yu Zhang, Yu Meng, Jiaxin Huang, Frank F Xu, Xuan Wang, and Jiawei Han. 2020. Minimally supervised categorization of text with metadata. *arXiv preprint arXiv:2005.00624*.

Yu Zhang, Wei Wei, Binxuan Huang, Kathleen M Carley, and Yan Zhang. 2017. Rate: Overcoming noise and sparsity of textual features in real-time location estimation. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 2423–2426.