Seed Word Selection for Weakly-Supervised Text Classification with Unsupervised Error Estimation

Yiping Jin[1,2], **Akshay Bhatia**[2], **Dittaya Wanvarie**[1]
[1]Department of Mathematics & Computer Science, Chulalongkorn University, Thailand
[2]Knorex, 140 Robinson Road, 14-16 Crown @ Robinson, Singapore
{jinyiping, akshay.bhatia}@knorex.com
Dittaya.W@chula.ac.th

Abstract

Weakly-supervised text classification aims to induce text classifiers from only a few user-provided seed words. The vast majority of previous work assumes high-quality seed words are given. However, the expert-annotated seed words are sometimes non-trivial to come up with. Furthermore, in the weakly-supervised learning setting, we do not have any labeled document to measure the seed words' efficacy, making the seed word selection process "a walk in the dark". In this work, we remove the need for expert-curated seed words by first mining (noisy) candidate seed words associated with the category names. We then train interim models with individual candidate seed words. Lastly, we estimate the interim models' error rate in an unsupervised manner. The seed words that yield the lowest estimated error rates are added to the final seed word set. A comprehensive evaluation of six binary classification tasks on four popular datasets demonstrates that the proposed method outperforms a baseline using only category name seed words and obtained comparable performance as a counterpart using expert-annotated seed words [1].

1 Introduction

Weakly-supervised text classification eliminates the need for any labeled document and induces classifiers with only a handful of carefully chosen seed words. However, some researchers pointed out that the choice of seed words has a significant impact on the performance of weakly-supervised models (Li et al., 2018; Jin et al., 2020). The vast majority of previous work assumed high-quality seed words are given. However, many seed words reported in previous work are not intuitive to come up with. For example, in Meng et al. (2019), the seed words used for the category "Soccer" are {cup, champions, united} instead of more intuitive keywords like "soccer" or "football". We conjecture the authors might have tried these more general keywords but avoided them because they do not perform well.

While it is common to use labeled corpora to evaluate weakly-supervised text classifiers in the literature, we do not have access to any labeled document for new categories in the real-world setting. Therefore, there is no way to measure the model's performance and select the seed words that yield the best accuracy. A similar concern on assessing active learning performance at runtime has been raised by Kottke et al. (2019).

In this work, we device *OptimSeed*, a novel framework to automatically compose and select seed words for weakly-supervised text classification. We firstly mine (noisy) candidate seed words associated with the category names. We then train interim models with individual candidate seed words in an iterative manner. Lastly, we use an unsupervised error estimation method to estimate the interim models' error rates. The keywords that yield the lowest estimated error rates are selected as the final seed word set. A comprehensive evaluation of six classification tasks on four popular datasets demonstrates the effectiveness of the proposed method. The proposed method outperforms a baseline using only the category name as seed word and obtained comparable performance as a counterpart using expert-annotated seed words. We use binary classification as a case study in this work, while the idea can be generalized to multi-class classification using one-vs.-rest strategy.

The contributions of this work are three-fold:

1. We propose a novel combination of unsupervised error estimation and weakly-supervised text classification to improve the classification performance and robustness.
2. We conduct an in-depth study on the impact of different seed words on weakly-supervised text classification, supported by experiments

[1]Source code can be found at https://github.com/YipingNUS/OptimSeed.

with various models and classification tasks.

3. The proposed method generates keyword sets that yield consistent and competitive performance against expert-curated seed words.

2 Related Work

We review the literature in three related fields: (1) weakly-supervised text classification, (2) unsupervised error estimation, and (3) keyword mining.

2.1 Weakly-Supervised Text Classification

Weakly-supervised text classification (Druck et al., 2008; Meng et al., 2018, 2019) aims to use a handful of labeled seed words to induce text classifiers instead of relying on labeled documents.

Druck et al. (2008) proposed generalized expectation (GE), which specifies the expected posterior probability of labeled seed words appearing in each category. GE is trained by optimizing towards satisfying the posterior constraints without making use of pseudo-labeled documents.

Chang et al. (2008) introduced the first embedding based weakly-supervised text classification method. They mapped category names and documents into the same semantic space. Document classification is then performed by searching for the nearest category embedding given an input document.

Meng et al. (2018) proposed weakly-supervised neural text classification. They generate unambiguous pseudo-documents, which are used to induce text classifiers with different architectures such as convolutional neural networks (Kim, 2014) or Hierarchical Attention Network (Yang et al., 2016).

Recently, Mekala and Shang (Mekala and Shang, 2020) disambiguate the seed words by explicitly learning different senses of each word with contextualized word embeddings. They first performed k-means clustering for each word in the vocabulary to identify potentially different senses, then eliminated the ambiguous keyword senses.

Two most recent works developed concurrently but independently from our work (Meng et al., 2020; Wang et al., 2020) addressed the same task we are tackling: weakly-supervised text classification from only the category name. They both tap on the category names' contextualized representation and expand the seed word list by finding other words that would fit into the same context.

2.2 Unsupervised Error Estimation

Unsupervised error estimation aims to estimate the error rate of a list of classifiers *without a labeled evaluation dataset*. It is widely relevant to machine learning models in production, such as when a pre-trained model is applied to a new domain or when labeled dataset is costly to obtain. To our best knowledge, no previous work in weakly-supervised classification applied unsupervised error estimation. Instead, they trained classifiers without labeled *training* datasets but evaluated their models used labeled *evaluation* datasets.

Most work in unsupervised error estimation derive the error rate analytically by making simplifying assumptions. Donmez et al. (2010) and Jaffe et al. (2015) assumed the marginal probability of the category $p(y)$ is known. Platanios et al. (2014) assumed classifiers make conditionally independent errors. While these approaches laid an important theoretical foundation, most assumptions cannot be met for real-world datasets and classifiers.

Platanios et al. (2016) proposed a Bayesian approach for error estimation. The model infers the true category and the error rates jointly using Gibbs sampling. The approach was benchmarked with various baselines such as majority vote and Platanios et al. (2014) and achieved superior performance. The estimated accuracy is usually within a few percents from the true accuracy. Notably, the only mild assumption it makes is that more than half of the classifiers have an error rate lower than 50%.

2.3 Keyword Mining

Keyword mining aims to bootstrap high-quality keyword lexicons from a small set of seed words, and it has been widely used in mining opinion lexicons (Hu and Liu, 2004; Hai et al., 2012) and technical glossaries (Elhadad and Sutaria, 2007). We want to draw the association between keyword mining and weakly-supervised text classification. Both tasks take a small list of seed words and unlabeled corpus as input, aiming to "expand" the knowledge about the target semantic category. Having more high-quality keywords will improve classification accuracy, while an accurate classifier will make the keyword mining task much easier by eliminating irrelevant/noisy documents.

3 Method

Figure 1 overviews OptimSeed, a framework to select seed words for weakly-supervised text classifi-

cation involving the following steps: (1) expanding candidate keywords from a single seed word, (2) training interim classifiers with individual candidate seed keywords using weakly supervision, (3) select the final seed words with the feedback from unsupervised error estimation. We discuss the proposed framework in detail in the following sections. To make our paper self-contained, we also brief the weakly-supervised classification and unsupervised error estimation model we use.

3.1 Expanding Candidate Keywords from a Single Seed

We use either the category name or trivial keywords (e.g., "good" and "bad" for sentiment classification tasks) as the only input seed word and use a keyword expansion algorithm to mine more candidate keywords. We apply $pmi\text{-}freq$ (Equation 1) following Jin et al. (2020). It is a product of the logarithm of the candidate keyword w's document frequency and the point-wise mutual information between w and the seed word s. The higher the $pmi\text{-}freq$ score, the more strongly the candidate keyword is associated with the seed word s. Additionally, we filter the mined keywords based on their part-of-speech tag depending on the classification task. We keep only noun candidates for topic classification and adjective candidates for sentiment classification.

$$pmi\text{-}freq(w;s) \equiv log\, df(w) log \frac{p(w,s)}{p(w)p(s)} \quad (1)$$

3.2 Training interim classifiers

The candidate keywords and unlabeled dataset are used to induce *interim classifiers*. Interim classifiers' purpose is to isolate the impact of individual seed words so that we can rank them. Specifically, iteration A in Figure 1 tries to rank candidate seed words for Category A (Movies) in the classification task Movies-Television. The initial seed word "television" for Category B is fixed, and it forms seed word tuples with each candidate word in Category A. We use each such seed word tuple as input to train an interim classifier. We then use each interim classifier's predictions to perform unsupervised error estimation.

We use Generalized Expectation (GE) (Druck et al., 2008) to train both interim classifiers and the final classifier because of its competitive performance and fast training speed [2]. GE translates labeled keywords to constraint functions. For example, the first keyword tuple (hollywood, television) in Figure 1 translates to two constraint functions: $hollywood \rightarrow A:0.9, B:0.1$ and $television \rightarrow A:0.1, B:0.9$, which means "hollywood" is expected to occur 90% in a document of category A while 10% in a document of category B, vice versa for the keyword "television".

Each constraint function on a labeled word w_k contributes to a term in the objective function in Equation 2 and the underlying logistic regression model is trained by minimizing the L2 distance between the reference distribution $\hat{p}(y|w_k > 0)$ (specified by the constraint function) and the empirical distribution $\tilde{p}(y|w_k > 0)$ (predicted by the model) of the category y when word w_k is present.

$$\mathcal{O} = -\sum_{k \in K} dist(\hat{p}(y|w_k > 0)||\tilde{p}(y|w_k > 0)) \quad (2)$$

3.3 Keyword Evaluation with Bayesian Error Estimation

We apply unsupervised error estimation on the interim classifiers' predictions to estimate their accuracy and select the best seed words for the final classifier. As shown in Figure 1 iteration A, the three keywords "hollywood", "filmmaker", and "theaters" are added to the final seed word set of Category A (Movies) because their corresponding interim classifiers have estimated accuracy above the threshold. The process is repeated in iteration B to select seed words for Category B.

We use the Bayesian error estimation (BEE) model (Platanios et al., 2016) to perform this step. In BEE, each instance's true label is latent, while each model's predictions are observed. The accuracy/error rate can be derived from the predictions and the latent true labels. The assumption that half of the classifiers have an error rate below 50% implicitly uses inter-classifier agreement.

BEE uses Gibbs sampling to infer the error rates of individual classifiers e_j and the true label l_i jointly. We refer the readers to Section 4.1 in Platanios et al. (2016) for the exact conditional probabilities used in Gibbs sampling.

[2] All GE models in this work can be trained within a few seconds using a single CPU core.

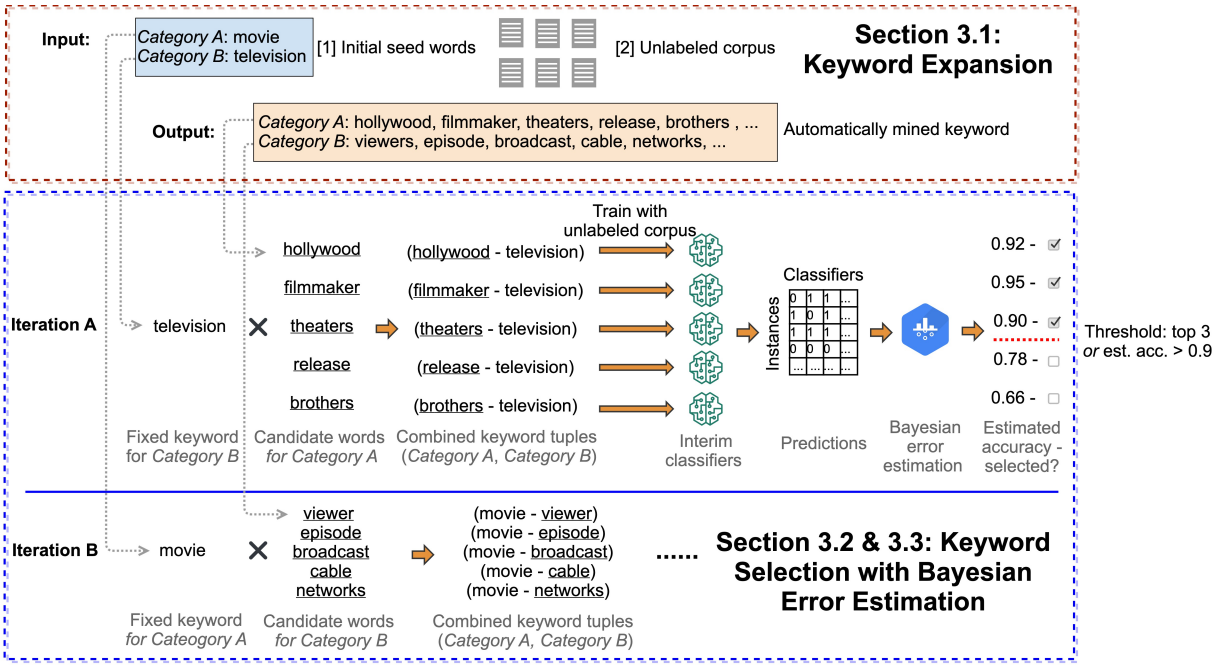

Figure 1: OptimSeed, a method to select seed words for weakly-supervised text classification. We first mine noisy keywords associated with the category name (the initial seed word). We use one iteration to refine the keywords for each category. In each iteration, We fix the seed word for one category and combine it with each mined keyword in the other category. The resultant keyword tuples are used to train separate interim classifiers. Finally, we use Bayesian error estimation to estimate the accuracy of classifiers induced from each keyword tuple and select the keywords with the highest estimated accuracy.

4 Experimental Setup

We use six binary classification tasks from four datasets to evaluate our framework. We choose the evaluation tasks so that they cover different granularities and domains. The details are as follows:

- **AG's News Dataset:** contains 120,000 documents evenly distributed into 4 coarse categories. We randomly choose two binary classification tasks: "Politics" vs. "Technology" and "Business" vs. "Sports".
- **The New York Times (NYT) Dataset:** contains 13,081 news articles covering 5 coarse and 25 fine-grained categories. We choose two fine-grained binary classification tasks involving categories with similar semantics: "International Business" (InterBiz) vs. "Economy" and "Movies" vs. "Television".
- **Yelp Restaurant Review Dataset:** contains 38,000 reviews evenly distributed into 2 categories: "Positive" vs. "Negative".
- **IMDB Movie Review Dataset:** contains 50,000 reviews evenly distributed into 2 categories: "Positive" vs. "Negative".

We report the performance of the following weakly-supervised models besides Generalized Expectation (GE):

- **Dataless (Chang et al., 2008):** maps both input documents and category seed words into a semantic space using Explicit Semantic Analysis (ESA) (Gabrilovich et al., 2007) over Wikipedia concepts and assigns the category nearest to the input document's embedding.
- **MNB/Priors (Settles, 2011):** increases priors for labeled keywords in a Naïve Bayes model and learns from an unlabeled corpus using EM algorithm.
- **WeSTClass (Meng et al., 2018):** weakly-supervised neural text classifier trained using pseudo documents. We use the CNN architecture because Meng et al. (2018) showed that it outperformed other architectures such as RNNs and Hierarchical Attention Network.
- **ConWea (Mekala and Shang, 2020):** leverages contextualized word representations to differentiate multiple senses. It also trains classifiers and expands seed words in an iterative manner.

We also report the performance of **LR**, a supervised logistic regression model trained using all the documents in the training set [3].

In all experiments, we mine 16 candidate seed words with the highest $pmi\text{-}freq$ score for each category. We select a candidate keyword for the final classifier if its estimated accuracy is among the top three or is higher than 0.9 [4]. For GE, we use a reference distribution of 0.9 (meaning a labeled keyword is expected to appear in its specified categories 90% of the time) following Druck et al. (2008). Table 1 shows the initial seed words used in our work and in previous work [5].

Class	Our Work	Previous Work
Politics	political;	democracy religion liberal;
Tech	technology	scientists biological computing
Business	business;	economy industry investment;
Sports	sports	hockey tennis basketball
InterBiz	international;	china union euro;
Economy	economy	fed economists economist
Movies	movie;	hollywood directed oscar;
Television	television	episode viewers episodes
Yelp & IMDB	good;	terrific great awesome;
	bad	horrible subpar disappointing

Table 1: Initial seed words for each task.

5 Classification Performance

Table 2 presents each model's average accuracy across six datasets.

We can see that OptimSeed seed words yield better performance than using the category name alone by a large margin for all weakly-supervised models,

Method	cate	ours	gold
Dataless	54.7	**60.4***	56.7
MNB/Priors	68.5	71.7	**74.4**
WeSTClass	75.7	**77.2**	77.0
ConWea	60.0	66.0	**70.7**
GE	80.4	84.8*	**85.1**
LR		91.8	

Table 2: Average accuracy scores in percentage for all methods on all six classification tasks. **cate**, **ours**, **gold** indicates the result using the category name, keywords selected by OptimSeed and expert-composed keywords used in previous work. For each model, the best-performing keyword set is highlighted in bold. * indicates statistical significance from the same model using "cate" seed word with p-value of 0.05 using paired t-test.

validating the effectiveness of our seed word expansion and selection method. It also achieved better or similar performance as expert-curated seed words for three out of five models.

Among the learning algorithms, GE obtained the best average performance for all seed word sets. The average accuracy of GE using OptimSeed seed words (84.8%) is only 0.3% lower than using expert-curated seed words, virtually eliminating human experts from the loop. GE+OptimSeed's accuracy is 7% below a fully-supervised logistic regression model trained on hundreds to tens of thousands of labeled documents.

Table 3 shows each model's classification accuracy on topic classification tasks. Summing over all models and datasets, OptimSeed achieved better or equal performance than the category name baseline 80% of the time (16/20) and better or equal performance than the gold seed words 65% of the time. It demonstrates the robustness of our seed word selection method.

While ConWea claimed to resolve ambiguity through contextualized embeddings, we observed that it works well only when the input seed words are unambiguous ("ours" or "gold" column). On the Business-Sports classification task, its accuracy was only 39.1% while other baselines could achieve over 90%. We inspected the model and found the keywords expanded by ConWea are much noisier than OptimSeed, which caused the poor performance.

We can make similar observations on the performance of sentiment classification tasks (Table 4). However, the gap between weakly-supervised mod-

[3] We use the logistic regression implementation in scikit-learn with default parameters and tf-idf features.

[4] Mekala and Shang (2020) observed that three seed words per class are needed for reasonable performance while more high-quality keywords help. Therefore, we use the accuracy threshold of 0.9 to include additional keywords.

[5] The seed words for NYT corpus were reported in Meng et al. (2019) and the rest are from Meng et al. (2018). No previous work in weakly supervision used IMDB dataset, so we use the same manual seed words as Yelp dataset.

Method	Poli-Tech			Biz-Sport			IB-Econ			Movie-TV		
	cate	ours	gold	cate	ours	gold	cate	ours	gold	cate	ours	gold
Dataless	50.1	**51.4**	50.2	50.0	50.2	**50.4**	59.1	**75.0**	67.1	67.8	**70.0**	67.8
MNB/Priors	87.3	**88.9**	**88.9**	**95.6**	93.9	92.9	58.5	54.3	**93.9**	67.8	67.8	**68.9**
WeSTClass	87.4	**89.5**	88.8	92.7	**94.8**	94.3	77.7	**83.0**	75.1	50.4	**76.6**	62.1
ConWea	71.5	**73.7**	71.4	39.1	67.0	**82.0**	75.1	71.2	**84.3**	66.9	**77.0**	76.4
GE	86.9	87.8	**88.5**	**93.0**	**93.0**	79.4	70.7	81.7	**91.5**	94.4	**98.9**	97.8
lr		96.3			98.6			90.2			85.5	

Table 3: Accuracy on topic classification tasks. For each model-dataset combination, we highlight the best performance in bold.

els and the supervised baseline is much larger, suggesting that some reviews' sentiment might be expressed implicitly and requires more than word-level understanding. Meng et al. (2020) also made a similar remark based on their experiment.

Method	Yelp		
	cate	ours	gold
Dataless	51.0	**55.5**	52.2
MNB/Priors	50.9	**71.5**	51.7
WeSTClass	78.3	58.8	**81.5**
ConWea	51.0	**51.3**	50.7
GE	68.0	75.2	**79.3**
lr		92.2	
Method	**IMDB**		
	cate	**ours**	**gold**
Dataless	50.1	**60.4**	52.2
MNB/Priors	51.1	**54.0**	50.3
WeSTClass	**67.7**	60.6	60.5
ConWea	56.5	55.7	**59.1**
GE	69.6	72.2	**74.0**
lr		88.3	

Table 4: Accuracy on sentiment classification tasks. For each model-dataset combination, we highlight the best performance in bold.

6 Case Study

To demonstrate the working of our proposed framework, we present a case study on the classification task "International Business" vs. "Economy" in Table 5 and show different seed word sets for the category "economy" and their corresponding performance.

Keyword expansion alone improved the accuracy significantly from the category name baseline. However, it may introduce some ambiguous keywords in the meantime. The unsupervised error estimator successfully identified top keywords such as "economist" and "economists" and eliminated poor keywords like "purchases" and "growth", which further improved the accuracy by 2.4%.

Stage:Acc	Seed Words for "Economy"
Init: 70.7	economy
Keyword Expansion: 79.3	purchases pace index borrowing unemployment economists economy stimulus rates recovery economist rate fed reserve inflation growth
Final: 81.7	economist economists rate recovery index

Table 5: Seed words for "Economy" at different stages of the OptimSeed framework.

7 Conclusion

Weakly-supervised text classification can induce classifiers with a handful of carefully-chosen seed words instead of labeled documents. However, the choice of seed words has a significant impact on classification performance. We proposed *OptimSeed*, a novel framework to compose the seed words automatically. It first mines keywords associated with the category name and then estimates each seed word's impact directly using unsupervised error estimation. The framework outputs seed words yielding a comparable performance as expert-curated seed words, virtually eliminating human experts from the loop.

8 Acknowledgements

YJ was supported by the scholarship from 'The 100th Anniversary Chulalongkorn University Fund for Doctoral Scholarship'. We thank anonymous reviewers for their valuable feedback.

References

Ming-Wei Chang, Lev-Arie Ratinov, Dan Roth, and Vivek Srikumar. 2008. Importance of semantic representation: Dataless classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 2, pages 830–835.

Pinar Donmez, Guy Lebanon, and Krishnakumar Balasubramanian. 2010. Unsupervised supervised learning i: Estimating classification and regression errors without labels. *Journal of Machine Learning Research*, 11(4).

Gregory Druck, Gideon Mann, and Andrew McCallum. 2008. Learning from labeled features using generalized expectation criteria. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 595–602. ACM.

Noemie Elhadad and Komal Sutaria. 2007. Mining a lexicon of technical terms and lay equivalents. In *Proceedings of Biological, Translational, and Clinical Language Processing*, pages 49–56, Prague, Czech Republic. Association for Computational Linguistics.

Evgeniy Gabrilovich, Shaul Markovitch, et al. 2007. Computing semantic relatedness using wikipedia-based explicit semantic analysis. In *Proceedings of the Twentieth International Joint Conference on Artificial Intelligence*, volume 7, pages 1606–1611.

Zhen Hai, Kuiyu Chang, and Gao Cong. 2012. One seed to find them all: mining opinion features via association. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, pages 255–264.

Minqing Hu and Bing Liu. 2004. Mining opinion features in customer reviews. In *Proceedings of the 19th National Conference on Artifical Intelligence*, pages 755–760.

Ariel Jaffe, Boaz Nadler, and Yuval Kluger. 2015. Estimating the accuracies of multiple classifiers without labeled data. In *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pages 407–415.

Yiping Jin, Dittaya Wanvarie, and Phu TV Le. 2020. Learning from noisy out-of-domain corpus using dataless classification. *Natural Language Engineering*, In press:1–35.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751.

Daniel Kottke, Jim Schellinger, Denis Huseljic, and Bernhard Sick. 2019. Limitations of assessing active learning performance at runtime. *arXiv preprint arXiv:1901.10338*.

Chenliang Li, Shiqian Chen, Jian Xing, Aixin Sun, and Zongyang Ma. 2018. Seed-guided topic model for document filtering and classification. *ACM Transactions on Information Systems (TOIS)*, 37(1):1–37.

Dheeraj Mekala and Jingbo Shang. 2020. Contextualized weak supervision for text classification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 323–333.

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.

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.

Yu Meng, Yunyi Zhang, Jiaxin Huang, Chenyan Xiong, Heng Ji, Chao Zhang, and Jiawei Han. 2020. Text classification using label names only: A language model self-training approach. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Emmanouil Antonios Platanios, Avrim Blum, and Tom Mitchell. 2014. Estimating accuracy from unlabeled data. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, pages 682–691, Arlington, Virginia, USA. AUAI Press.

Emmanouil Antonios Platanios, Avinava Dubey, and Tom Mitchell. 2016. Estimating accuracy from unlabeled data: A bayesian approach. In *Proceedings of the International Conference on Machine Learning*, pages 1416–1425.

Burr Settles. 2011. Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1467–1478. Association for Computational Linguistics.

Zihan Wang, Dheeraj Mekala, and Jingbo Shang. 2020. X-class: Text classification with extremely weak supervision.

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.