Incomplete Supervision: Text Classification based on a Subset of Labels

Luning Yang
l4yang@ucsd.edu

Yacun Wang
yaw006@ucsd.edu

Abstract

Many text classification models rely on an assumption that requires users to provide the model with a full set of class labels. This is not realistic, as users may not be aware of all possible classes in advance, or may not be able to obtain an exhaustive list. Thus, we propose to work in a new setting where both labeled and unlabeled articles exist, and aim to discover classes among the unlabeled articles. We explore the potential of weakly supervised ML to detect class labels that humans may not recognize, thus facilitating more accurate classification. At this time, the baseline model learns well from the supervised set, but label generation is less satisfactory.

1 Introduction

In recent years, with the growing complexity and scale of neural network models, they also require more high-quality human-annotated training data to achieve satisfactory performances. These actions usually require extensive domain expertise and are extremely time-consuming. Researchers have strived to develop models in the weak supervision setting that aim to gradually alleviate the human burden in creating such annotations for the documents. In particular, researchers have approached the problem of text classification by developing models that only require the class labels and a little extra information for each class label such as (1) a few representative words (i.e. seed words); (2) authors, publication date, etc. (i.e. metadata). Researchers have shown that models are capable of obtaining reliable results without full human annotation.

However, the problem setting for these models all depend on one key assumption: users need to provide the model with a full set of desired class labels for the model to consider. This is less realistic as users might not know all possible classes in advance; users are also unable to obtain an exhaustive list of class names without carefully reading and analyzing the documents.

For example, an online article database might contain thousands of user-uploaded articles with only a subset labeled with their domains such as news, sports, or computer science. The remaining articles are left unlabeled. However, we have to admit that the labels provided by the database are limited. There might be some classes existing in our documents whose labels are not provided by our database. For instance, we may have a group of articles in the domain of chemistry, while we don't have the exact label "chemistry" for that group.

In this paper, we work in a new setting where both labeled and unlabeled articles exist. We aim to assign accurate existing labels and also discover new classes among the unlabeled articles that have better choices. We try to explore the possibility of utilizing the power of machines to detect class labels that humans fail to recognize and classify documents to more reasonable labels. In particular, we proposed a baseline model and an advanced model that both leverage semi-supervised and unsupervised learning methods to extract information from the labeled part of the dataset, learn patterns from the unlabeled part, and generate new labels based on documents that have lower similarity between their representation and existing class labels. At this time, our baseline model is performing relatively well given the simplicity of sub-models chosen, but the labels generated are dominated by popular classes and the confident documents share a close-to-uniform representation distribution.

Table 1: Dataset Statistics

Dataset	# Docs	# Classes	Avg # Words
NYT-Fine	11,527	26	648.24
Reddit	48,407	20	24.11
DBPedia	560,000	14	50.01

2 Data

2.1 Datasets

We picked data from 3 categories: news, social media, and Wikipedia. The basic statistics are shown in 1.

- **The New York Times (nyt-fine):** The NYT dataset consists of news articles published by

the New York Times, and is reused from the ConWea paper (Mekala and Shang, 2020). There are 26 fine-grained categories which are stemmed from coarse grained labels (omitted), and the number of documents follow a long-tailed distribution.

- **DBPedia:** The articles come from topic classifications based on Wikipedia pages, and is reused from LOTClass (Meng et al., 2020) and X-Class (Wang et al., 2021). There are 14 perfectly balanced classes and a large number of documents.

- **Reddit:** The Reddit dataset contains social media posts from Reddit, which includes the post titles and descriptions. There are 20 classes following a long-tailed distribution.

2.2 Label Removal

We obtain a fully labelled dataset and remove part of the labels to conform with our task setting. We first obtain n documents $\{D_1, D_2, \ldots, D_n\}$ which are each labelled with one of the classes $c_1, \ldots, c_m$, and let f be a mapping from the documents to the labels. We assume the frequency of class labels follow a long-tailed distribution, for example Zipf's Law. Let the frequency of the labels be $f_i = \#\{k : f(D_k) = i\}$, where the labels are ranked by their frequency in descending order.

In the incomplete setting, we also assume: (1) new labels failed to be provided by users all come from less frequent classes; (2) the less frequent, the more likely that it's missed from the users. To create datasets of such a setting, we sample labels from the less frequent half:

From labels $c_{m/2+1}, \ldots, c_m$, we sample l labels from the discrete distribution for each i in the bottom half:

$$P(X = i) = \frac{1/f(i)}{\sum_{j=m/2+1}^{m} 1/f(j)}$$

Finally, from the remaining labels, we remove the same percentage p of labels from documents to gain our final data set.

3 Problem

Extensive research has been conducted to classify documents using a decreasing amount of human effort. Meng et al., 2018; Mekala and Shang, 2020 utilize a few user-provided seed words for the only annotation; Mekala et al., 2020; Zhang et al., 2022 utilize metadata information about the articles such as authors and publication dates which require less human effort to obtain; more advanced methods such as Meng et al., 2020; Wang et al., 2021; Zeng et al., 2022; Shen et al., 2021 only require class label names. All of the models leverage the limited information provided by humans with learned representations including knowledge graphs, large pre-trained language models, etc. to achieve satisfactory classification accuracy.

We attempt to further reduce user workload and perform text classification under the more new use case where the input class label set is only a proper subset of all possible labels being predicted. We define this setting to be the "incompletely" supervised text classification. This setting is more realistic because in most use cases users wish to classify their own unannotated corpus with some expectation and some labeled data, and the end task is to complete the labels for the entire corpus. Almost none of the previous research has mentioned this setting, including papers we have covered in Quarter 1, but similar methods could be tweaked in minor ways and applied to this new setting.

We present the full problem statement. The "incomplete" supervision takes in a set of documents $D = \{D_1, \ldots, D_n\}$, a user-provided set of desired class labels $c^* = \{c_1, \ldots, c_{m^*}\}$, and a set of labels for the first k documents $l_1, \ldots, l_k \in c^*$ The task is to suggest new possible class names $c_{m^*+1}, \ldots, c_m \notin c^*$ to form the full label set $c = \{c_1, \ldots, c_m\}$ and assign a reasonable label $l_j \in c$ for $j \in k+1, \ldots, n$ for the remaining unlabeled documents. In total, we will predict one of the m labels for $n-k$ documents given m^* original labels and k labeled documents.

4 Evaluation

Since the model generates newly suggested labels that are likely to not exist in the original set of full labels, we design the following evaluation process to assess the quality of newly generated labels as well as aggregated classification performance. The term "existing label" refers to the ground truth for documents that exist in the labeled part, and thus will exist in the unlabeled part; in contrary, "newly generated label" refers to ground truths that only exist in the unlabeled part (i.e. is removed during the label removal process).

As described in the models, we allow each docu-

ment to first select from the existing labels before moving towards generating new labels. Our evaluation process follow a similar multi-step framework to aim on different parts of the model.

4.1 Binary Classification

The model decides on whether to generate new labels for a document based on the confidence of weak supervised document-class representations. The sub-task of predicting whether a document falls outside of existing classes is a binary classification prediction. We evaluate this sub-task using binary precision and recall, with "new labels necessary" as the positive class.

4.2 Existing Label Performance

Based on all documents that have ground truth as existing labels, we evaluate the multi-class classification using the micro- and macro-F1 scores.

4.3 New Label Inspection

After new labels are generated, we may inspect the quality of new labels using either manual inspection, or automatically using pre-trained Word2Vec models to map each generated label to the most similar removed label.

4.4 Automatic Full Evaluation

With the second automatic mapping method described in the previous section, we might construct the full set of predictions using existing labels predictions and mapped new label generation. We report both the micro- and macro-F1 scores.

5 Baseline Model

5.1 Method

The models for the incomplete setting mainly contains 3 modules: (1) a supervised model learning useful information from the labeled set; (2) a weakly supervised model learning patterns of the unlabeled set using the information; (3) an unsupervised model clustering documents and generating new labels.

The baseline model is a combination of a few vanilla basic models: (1) a supervised TF-IDF model for seed word learning; (2) a weakly supervised Word2Vec model that takes in seed words and output document and class representations; (3) a cosine similarity measure for confidence split; (4) a Gaussian Mixture Model for clustering unconfident documents; (5) another unsupervised TF-IDF module for label generation from the clusters. Figure 1 illustrates the steps taken in the baseline model.

The purpose of each sub-model is as follows:

First, we extract the top 10 unique words per label from TF-IDF scores and create a seedword set. To ensure the quality and accuracy of the seedword sets, if two labels share a common seedword, it is removed and replaced with the label's following most frequent words.

Once the seedword set for each label is finalized, a Word2Vec (Mikolov et al., 2013) model is trained to learn word embedding vectors for each word in the unlabeled corpus. To predict the label, we separately find the document and label representations by aggregating the vector representations:

$$v_l = \frac{1}{|S_l|} \sum_{s \in S_l} v_s, \quad v_d = \frac{1}{|W_d|} \sum_{w \in W_d} v_w$$

where S_l is the seed word set for label l, W_d is the words in document d.

Then we find the relevance by taking the cosine similarity between documents and labels. In particular, we to predict the label $\hat{l}$ for document d, we find:

$$\hat{l}_d = \underset{l \in L}{\operatorname{argmax}} \frac{\langle v_l, v_d \rangle}{\|v_l\| \cdot \|v_d\|}$$

where L is the set of labels.

To split unconfident documents for new label generation, we create a new set of documents whose cosine similarity score falls below the threshold τ. These removed texts are assumed to be relatively irrelevant to the set of existing labels. The removed documents, along with their document representation v_d, are taken together to run a Gaussian Mixture clustering algorithm.

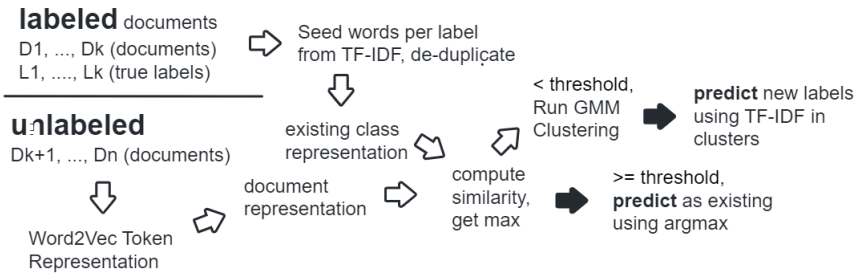
Gaussian Mixture is a probabilistic model-based soft clustering algorithm. For our representations $V = \{v_1, \ldots, v_n\}$ of the n unconfident documents, the objective function to maximize is:

$$P(V|C) = \prod_{i=1}^{n} P(v_i|C) = \prod_{i=1}^{n} \sum_{j=1}^{K} w_i f_j(v_i)$$

where $C = \{C_j\}_{j=1}^{K}$ is the set of clusters, where each cluster has a Gaussian density function f_j and a prior distribution w_j.

The classic way is to find the maximum likelihood estimation using the Expectation-Maximization (EM) approach. We iteratively repeat the expectation and maximization steps until

Figure 1: Baseline Model Illustration



convergence. In the expectation (E) step, we assign text embedding vectors to clusters according to the current parameters of probabilistic clusters. In the maximization (M) step, we find the new clustering or parameters that minimize the sum of squares error (SSE) or the expected likelihood. Finally, we assign each text embedding vector to the cluster with the highest weight.

The clusters from the GMM clustering helps us find the new classes, and we generate the new label using the unique word with the best TF-IDF score for each class.

5.2 Experiment Settings

For the baseline model, we mainly tested on the New York Times fine-grained dataset. There is much room for hyperparameter tuning, but learning from the ConWea replication results as well as for the purpose of simplicity, the following choices are made:

- **Preprocessing:** Remove punctuation and standard English stopwords;
- **Word2Vec:** Vector Dimension 128, Window 10, Epochs 150;
- **Confidence Split**: Threshold 0.25

The remaining knobs include the unconfident split threshold for new label generation based on the distribution of maximum cosine similarity, as well as the number of classes for the clustering algorithm to predict.

5.3 Results and Discussion

We report the results of the baseline model for each step described above, if applicable.

Seed Words: We present a few example seed words generated from the first supervised TF-IDF module below. The basic TF-IDF scores are able to identify relatively representative seed words.

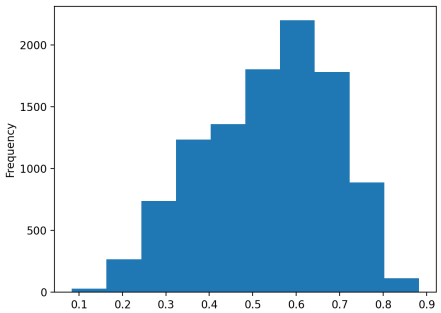


Figure 2: Maximum Similarity Distribution for Unlabeled Documents

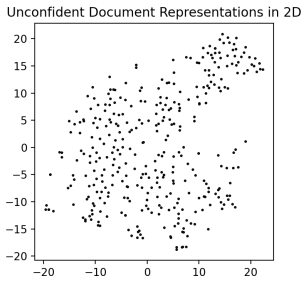


Figure 3: t-SNE Dimensionality Reduction for Unlabeled Documents

- baseball: inning, yankees, hit, ...
- basketball: knicks, nets, rebounds, ...
- cosmos: earth, bigelow, asteroid, ...

Similarity Distribution: Figure 2 shows the distribution of the maximum cosine similarity found for all unlabeled documents, and thus provides us with the criteria to get unconfident documents. From the figure, the distribution is roughly normal with a slight left skew; however, the lowest maximum similarity, around 0.1, is higher than expected. From the metric, we may expect the split to be less ideal.

Unconfident Documents: Figure 3 shows the 2D unconfident document representations after applying the t-SNE (van der Maaten and Hinton,

Table 2: Baseline model experiment results

Threshold	# Unconfident	New Label Binary		Existing Labels		New Labels (5 Clusters)
		Precision	Recall	Micro-F1	Macro-F1	
0.2	90	0.0	0.0	0.884	0.819	['braun', 'bats', 'twitter', 'nfl', 'hall']
0.25	354	0.014	0.016	0.886	0.824	['ncaa', 'braun', 'im', 'teams', 'mets']
0.3	759	0.024	0.058	0.88	0.824	['jeter', 'im', 'points', 'nfl', 'braun']
0.35	1423	0.039	0.179	0.86	0.821	['yankees', 'points', 'mr', 'cup', 'coach']
ConWea Replication: Best Word2Vec				0.75	0.63	-

2008) dimensionality reduction technique to visualize the high-dimensional data. To generate the 2D representation, we followed the suggestions on `sklearn` t-SNE to first use Principle Component Analysis (PCA) to reduce to 50 dimensions, and apply t-SNE with perplexity 30. From the figure and some other perplexity experiments, the distribution of representations are roughly uniform, indicating that the unconfident documents have little unique pattern. Therefore, we expect the new label generation to be hard from this result as well.

Experiment Results: Table 2 shows the results of experimenting with different threshold cutoffs. From the existing labels prediction, we could see that the supervised + weakly supervised models perform relatively well on classes already known in the dataset. Specifically, compared to ConWea replication where seed words are all human-chosen, the ability for the model to learn the seed words from the existing labeled documents are helping the understanding of existing classes. Note that sometimes in the ConWea setting we might not have enough labeled documents to generate the seed words, so human effort is still useful.

In contrast, the new label binary classification gives almost the worst result, providing near 0 precision and recall of detecting documents in need of new labels. This indicates that the good results from existing labels are very likely to be dominated by popular classes. These classes are also popular in the labeled part, it's likely that the model learn better representative seed words. However, because of the same reason, minority classes are underrepresented and the model has almost no ability of discerning between documents belonging to existing or new classes. The values do increase with increasing threshold, but that's mostly due to more documents are being considered as unconfident. Another reason relates to the NYT dataset – as the labels are fine grained, it's likely for the model to map to other classes in the same coarse label, and thus are unable to identify new classes. The results from the similarity and t-SNE plots also support this finding.

With the low binary classification results, it's no surprise that the labels generated are not close to the truly removed labels, but we could detect a pattern of increasing dominance of popular classes in the newly generated labels. When the threshold is increased, the more documents from existing classes and thus popular classes become part of the unconfident set, and the new labels detected by TF-IDF will be closer to represent popular classes.

6 Advanced Model

With the shortcomings detected in the baseline model, the advanced model will focus on improving the ability to distinguish existing and new classes.

6.1 Method

The methods are still under construction.

6.2 Experiment Settings

The experiment settings are still under construction.

6.3 Results and Discussion

The results and discussion are still under construction.

7 Conclusion

In conclusion, the setting of incomplete text classification needs specialized methods to deal with less popular classes, as documents in need of new labels are mostly in the long-tailed distribution. If the confidence split is unable to produce documents with specialized patterns, the label generation process will not give promising results.

Contributions

Both members contributed extensively to the construction, brainstorm, implementation, and delivery of this project.

Luning Yang participated in the brainstorming discussion, set up the reddit API server and collected data from selected subreddits, drafted the baseline model code with main ideas on detailed model choices, and drafted part of the report.

Yacun Wang participated in the brainstorming discussion, pre-processed the raw data, converted the dataset into the incomplete supervised setting, revised and finalized the baseline model code, configured the GitHub submission, and drafted part of the report.

References

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, Online. Association for Computational Linguistics.

Dheeraj Mekala, Xinyang Zhang, and Jingbo Shang. 2020. META: Metadata-empowered weak supervision for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8351–8361, Online. Association for Computational Linguistics.

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*, CIKM '18, page 983–992, New York, NY, USA. Association for Computing Machinery.

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 (EMNLP)*, pages 9006–9017, Online. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Jiaming Shen, Wenda Qiu, Yu Meng, Jingbo Shang, Xiang Ren, and Jiawei Han. 2021. TaxoClass: Hierarchical multi-label text classification using only class names. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4239–4249, Online. Association for Computational Linguistics.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Zihan Wang, Dheeraj Mekala, and Jingbo Shang. 2021. X-class: Text classification with extremely weak supervision. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Ziqian Zeng, Weimin Ni, Tianqing Fang, Xiang Li, Xinran Zhao, and Yangqiu Song. 2022. Weakly supervised text classification using supervision signals from a language model. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2295–2305, Seattle, United States. Association for Computational Linguistics.

Yu Zhang, Shweta Garg, Yu Meng, Xiusi Chen, and Jiawei Han. 2022. Motifclass: Weakly supervised text classification with higher-order metadata information. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, page 1357–1367, New York, NY, USA. Association for Computing Machinery.

Appendix: Proposal

1 Background

In recent years, with the growing complexity and scale of neural network models, they also require more high-quality human-annotated training data to achieve satisfactory performances. These actions usually require extensive domain expertise and are extremely time-consuming. Researchers have strived to develop models in the weak supervision setting that aim to gradually alleviate the human burden in creating such annotations for the documents. In particular, researchers have approached the problem of text classification by developing models that only require the class labels and a little extra information for each class label such as (1) a few representative words (i.e. seed words); (2) authors, publication date, etc. (i.e. metadata). Researchers have shown that models are capable of obtaining reliable results without full human annotation.

However, the problem setting for these models all depend on one key assumption: users need to provide the model with a full set of desired class labels for the model to consider. This is less realistic as users might not know all possible classes in advance; users are also unable to obtain an exhaustive list of class names without carefully reading and analyzing the documents.

For example, an online article database might contain thousands of user-uploaded articles with only a subset labeled with their domains such as news, sports, or computer science. The remaining articles are left unlabeled. However, we have to admit that the labels provided by the database are limited. There might be some classes existing in our documents whose labels are not provided by our database. For instance, we may have a group of articles in the domain of chemistry, while we don't have the exact label "chemistry" for that group.

Therefore, we wish to work in a new setting where both labeled and unlabeled articles exist. We aim to discover new classes among those unlabeled articles. We try to explore the possibility of utilizing the power of machines to detect class labels that humans fail to recognize and classify documents to more reasonable labels.

2 Problem

Extensive research has been conducted to classify documents using a decreasing amount of human effort. Meng et al., 2018; Mekala and Shang, 2020 utilize a few user-provided seed words for the only annotation; Mekala et al., 2020; Zhang et al., 2022 utilize metadata information about the articles such as authors and publication dates which require less human effort to obtain; more advanced methods such as Meng et al., 2020; Wang et al., 2021; Zeng et al., 2022; Shen et al., 2021 only require class label names. All of the models leverage the limited information provided by humans with learned representations including knowledge graphs, large pre-trained language models, etc. to achieve satisfactory classification accuracy.

We attempt to further reduce user workload and perform text classification under the more new use case where the input class label set is only a proper subset of all possible labels being predicted. We define this setting to be the "incompletely" supervised text classification. This setting is more realistic because in most use cases users wish to classify their own unannotated corpus with some expectation and some labeled data, and the end task is to complete the labels for the entire corpus. Almost none of the previous research has mentioned this setting, including papers we have covered in Quarter 1, but similar methods could be tweaked in minor ways and applied to this new setting.

We present the full problem statement. The "incomplete" supervision takes in a set of documents $D = \{D_1, \ldots, D_n\}$, a user-provided set of desired class labels $c^* = \{c_1, \ldots, c_{m^*}\}$, and a set of labels for the first k documents $l_1, \ldots, l_k \in c^*$ The task is to suggest new possible class names $c_{m^*+1}, \ldots, c_m \notin c^*$ to form the full label set $c = \{c_1, \ldots, c_m\}$ and assign a reasonable label $l_j \in c$ for $j \in k+1, \ldots, n$ for the remaining unlabeled documents. In total, we will predict one of the m labels for $n-k$ documents given m^* original labels and k labeled documents.

Part of this project will be directly comparable to the Quarter 1 project because we will provide baseline models that set the playground for more advanced methods. Since we have provided our tuned results on the baseline methods for a more annotation-dependent setting (i.e. weak supervision with class labels and user-provided sets of seed words), we will investigate the differences between the two settings and how well baseline models address settings with different amounts of supervision.

3 Data

3.1 Preprocessing

We denote the universal set of classes in each of our dataset as X. We select a tiny subset of class Y. For the posts belonging to Y, we replace ALL their class names with nulls. For the articles in the rest of the classes $X - Y$, we replace SOME of their class names with nulls. From the above preprocessing, we will get the dataset which resembles the article database which we mentioned in the Background section.

3.2 Datasets

We aim to be comprehensive and use datasets from as many domains as possible to evaluate our models.

- **The New York Times (nyt-fine):** The NYT dataset consists of news articles published by the New York Times, and is reused from the ConWea paper (Mekala and Shang, 2020). There are 26 fine-grained categories which are stemmed from coarse grained labels (omitted), and the number of documents follow a long-tailed distribution.

- **DBPedia:** The articles come from topic classifications based on Wikipedia pages, and is reused from LOTClass (Meng et al., 2020) and X-Class (Wang et al., 2021). There are 14 perfectly balanced classes and a large number of documents.

- **Reddit:** The Reddit dataset contains social media posts from Reddit, which includes the post titles and descriptions. There are 20 classes following a long-tailed distribution.

3.3 Dataset Quality

At this moment we are not worried about the quality of dataset, because of the following reasons:

1. We have a handful of datasets both scraped from social media and pre-existing datasets such as news, Wikipedia, Yelp, etc. that have been compiled and accessible thanks to previous researchers;

2. We'll be able to report performances on datasets that have relatively low quality, because the requirements of the task setting could most likely be achieved.

4 Primary Output

The main purpose of this project is to investigate the new setting and research potential models, report on their performances, and discuss the advantages as well as limitations of the setting. We will be creating a report, if not a scientific paper for the final output. In the report, we will be presenting:

1. An evaluation method for the "incomplete" setting. Since our project involves a new module of suggesting new class labels, simply using macro and micro-F1 scores might not be the best way to evaluate the model performance, as the incorrectly suggested labels could ruin the metric;

2. Baseline method description and model performance on collected data, along with discussions of the result;

3. If time permits, advanced method description and model performance on collected data, along with discussions of the result;

4. If time permits, possible extensions of the model to multi-label tagging prediction, methods to maintain granularity of suggested labels, setting when no labeled documents are provided, etc.;

5. As the capstone project requires, a website showcasing the project details.

5 Primary Method Ideas

From the discussions of the (extremely) weak supervision papers, we have currently briefly touched on the following ideas:

1. Baseline: Counting methods. Propose new class names based on the terms that have high term frequency and high document frequency;

2. Baseline: Simple word representation. Use context-free Word2Vec (Mikolov et al., 2013) word representations, gain document and existing label representations, find documents that have cosine similarity lower than a threshold, and make clusters based on the "isolated" documents to suggest new class labels;

3. Advanced Idea Piece: Pre-trained language models. Gain static and contextualized representations from large language models such as BERT, GPT, etc.;

4. Advanced Idea Piece: Clustering. Use different clustering methods when needed, suggested are Gaussian Mixture Model (GMM), DBSCAN;

5. Advanced Idea Piece: Training supervised neural network model at the last step to obtain a classifier.

We will also be reading papers that are related to the weak supervision, extremely weak supervision, metadata + weak supervision, semi-supervised settings and gain new inspirations from the methods that have been highlighted by previous researchers.

References

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, Online. Association for Computational Linguistics.

Dheeraj Mekala, Xinyang Zhang, and Jingbo Shang. 2020. META: Metadata-empowered weak supervision for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8351–8361, Online. Association for Computational Linguistics.

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*, CIKM '18, page 983–992, New York, NY, USA. Association for Computing Machinery.

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 (EMNLP)*, pages 9006–9017, Online. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Jiaming Shen, Wenda Qiu, Yu Meng, Jingbo Shang, Xiang Ren, and Jiawei Han. 2021. TaxoClass: Hierarchical multi-label text classification using only class names. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4239–4249, Online. Association for Computational Linguistics.

Zihan Wang, Dheeraj Mekala, and Jingbo Shang. 2021. X-class: Text classification with extremely weak supervision. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Ziqian Zeng, Weimin Ni, Tianqing Fang, Xiang Li, Xinran Zhao, and Yangqiu Song. 2022. Weakly supervised text classification using supervision signals from a language model. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2295–2305, Seattle, United States. Association for Computational Linguistics.

Yu Zhang, Shweta Garg, Yu Meng, Xiusi Chen, and Jiawei Han. 2022. Motifclass: Weakly supervised text classification with higher-order metadata information. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, WSDM '22, page 1357–1367, New York, NY, USA. Association for Computing Machinery.