LIME: Weakly-Supervised Text Classification Without Seeds

Seongmin Park and **Jihwa Lee**
ActionPower, Seoul, Republic of Korea
{seongmin.park, jihwa.lee}@actionpower.kr

Abstract

In weakly-supervised text classification, only label names act as sources of supervision. Predominant approaches to weakly-supervised text classification utilize a two-phase framework, where test samples are first assigned pseudo-labels and are then used to train a neural text classifier. In most previous work, the pseudo-labeling step is dependent on obtaining seed words that best capture the relevance of each class label. We present LIME[1], a framework for weakly-supervised text classification that entirely replaces the brittle seed-word generation process with entailment-based pseudo-classification. We find that combining weakly-supervised classification and textual entailment mitigates shortcomings of both, resulting in a more streamlined and effective classification pipeline. With just an off-the-shelf textual entailment model, LIME outperforms recent baselines in weakly-supervised text classification and achieves state-of-the-art in 4 benchmarks.

1 Introduction

Weakly-supervised text classification (Meng et al., 2018) is an important avenue of research in low-resourced text classification. Unlike in traditional text classification, all supervision derives from textual information in category names. Weakly-supervised classification offers a practical approach to classification because it does not necessitate massive amounts of training data.

Another distinct aspect of weakly-supervised text classification is that the system has access to the entire test set at evaluation time, instead of encountering test samples sequentially. Exploiting this characteristic, recent approaches employ keyword-matching pseudo-labeling schemes to tentatively assign class labels to each test sample, before using the information to train a separate classifier (Meng et al., 2018; Mekala and Shang, 2020; Wang et al., 2021). Pseudo-labels are assigned by counting how many "seed words" of each class are found in the test sample. Keyword matching-based labeling, however, is neither adaptable nor flexible because semantic information embedded in class names cannot be extracted adaptively for distinct classification tasks.

Inspired by recent advances in prompt-based text classification (Yin et al., 2019, 2020; Schick and Schütze, 2021), we replace the keyword-based pseudo-labeling step with a more streamlined entailment-based approach. Extensive experiments show that entailment-based classifiers assign more accurate pseudo-labels with greater task adaptability and much fewer hyperparameters. We find that our method realizes the benefits of both entailment-based classification and self-training.

Our contributions are as follows:

1. We present LIME, a novel framework for weakly-supervised text classification that utilizes textual entailment. LIME surpasses current state-of-the-art weakly-supervised methods in all tested benchmarks.

2. We show that self-training with pseudo-labels can mitigate unsolved robustness issues in entailment-based classification (Ma et al., 2021).

3. We experimentally confirm that higher confidence in pseudo-labels translates to better classification accuracy in self-training. We also find that a balance between filtering out low-confidence labels and preserving a sizable pseudo-training corpus is important.

2 Background

2.1 Weakly-supervised text classification

In weakly-supervised text classification, the system is allowed to view the entire test set at evaluation

[1] **L**abels **I**dentified with **M**aximal **E**ntailment

Proceedings of the 29th International Conference on Computational Linguistics, pages 1083–1088
October 12–17, 2022.

time. Having access to all test data allows novel preprocessing approaches unavailable in traditional text classification, such as preliminary clustering of test samples (Mekala and Shang, 2020; Wang et al., 2021) before attempting final classification. In the process, the system has an opportunity to examine overall characteristics of the test set.

Existing methods for weakly-supervised text classification focus on effectively leveraging such additional information. The dominant approach involves generating pseudo-data to train a neural text classifier. Most methods assign labels to samples in the test set by identifying operative keywords within the text (Meng et al., 2018). They obtain seed words that best represent each category name. Then, each sample in the test set is assigned a label with keywords most relevant to its content.

Later works improve this pipeline by automatically generating seed words (Meng et al., 2020b) or incorporating pre-trained language models to utilize contextual information of representative keywords (Mekala and Shang, 2020).

Seed-word-based pseudo-labeling, however, is heavily dependent on the existence of representative seed words in test samples. Seed-word-based matching cannot fully utilize information in contextual language representations, because the classification of each document involves brittle global hyperparameters such as the number of total seed words (Meng et al., 2020b) or word embedding distance (Wang et al., 2021).

In this work, we entirely forgo the seed word generation process during pseudo-labeling. We show that replacing seed-word generation with entailment-based text classification is more reliable and performant for text classification with weak supervision.

2.2 Entailment based text classification

Textual entailment (Fyodorov et al., 2000; MacCartney and Manning, 2009) measures the likeliness of a sentence appearing after another. Since entailment is evaluated to a probability value, the task can be extended for use in text classification. In entailment-based text classification, classification is posed as a textual entailment problem: given a test document, the system ranks the probabilities that sentences each containing a possible class label (*hypotheses*) will immediately follow the document text. The class label belonging to the most probable hypothesis is selected as the classification prediction. A hypothesis for topic classification, for example, could be "This text is about <*topic*>". The flexibility in prompt choices for constructing the hypotheses makes entailment-based classification extremely adaptable to different task types.

Although entailment-based sentence scoring is popular in zero- and few-shot text classification (Yin et al., 2019, 2020), the robustness of such approaches has recently been called into question (Ma et al., 2021). Since entailment-based classifiers rely heavily on lexical patterns, a large variance is observed in classification performance across different domains. We find that self-training commonly found in weakly-supervised classification mitigates such robustness issues in entailment-based classification to a large degree.

3 The LIME Framework

LIME enhances the two-phase weakly-supervised classification pipeline with an entailment-based pseudo-labeling scheme.

	Examples
Test sample (t)	"I love the food."
Class label (c)	"Positive"
Verbalizer	"Positive" $\rightarrow$ "*good*"
Prompt	"It was <$verbalizer(h_i)$>."
Hypothesis (h)	"It was good."

Table 1: Example test sample, class label, verbalizer, prompt, and entailment hypothesis. Converting class labels with a verbalizer is an optional procedure.

3.1 Phase 1: Pseudo-labeling

Textual entailment evaluates the likeliness of a hypothesis h succeeding some text t.

Given $C = \{c_1, c_2, \ldots, c_n\}$, the set of all possible labels for t, we generate $H = \{h_1, h_2, \ldots, h_n\}$, the set of all entailment hypothesis. Every sentence h_i asserts that its corresponding $c_i \in C$ is the correct label for t. h_i is constructed from a designated *prompt* and an optional *verbalizer* for each dataset (Schick and Schütze, 2021):

$$h_i = prompt(verbalizer(c_i))$$

Prompts dictate the wording of the hypotheses, while *verbalizers* convert each class label into a terminology better interpreted by entailment models. Pseudo-label for t is chosen as c_i that corresponds

Dataset	Type	# of Classes	Dataset size	Prompt
20News	News topic	5	17,871	*The text is about <class label>.*
AGNews	News topic	4	120,000	*The text is about <class label>.*
Yelp	Restarant review	2	38,000	*It was good. / It was bad.*
DBpedia	Wikipedia topic	14	560,000	*The text is about <class label>.*

Table 2: Statistics for benchmark datasets.

to the pair (t, h_i) with the highest entailment probability. Table 1 provides examples of verbalizers, prompts, and hypotheses.

3.2 Phase 2: Self-training

We adopt a similar self-training approach as existing methods in weakly-supervised text classification. We train a BERT-base model (Devlin et al., 2019) with a sequence classification feed-forward layer using pseudo-labels obtained in Phase 1.

We calculate the prediction confidence for each pseudo-label c_i assigned to t. Pseudo-labels under a certain confidence threshold are discarded during the text classifier training phase.

Confidence of label c_i is defined as the softmax over entailment probabilities of all hypotheses:

$$Confidence(c_i) = \frac{e^{p_i}}{\sum_{j=1}^{n} e^{p_j}}$$

where p_i is the entailment probability for the text pair (t, h_i), obtained from the entailment model.

4 Experiments

4.1 Experimental setting

In every experiment, we use a publically available BART-large model[2] (Lewis et al., 2020) trained on the MultiNLI (Williams et al., 2018) dataset as our entailment classifier. We also discard pseudo-labels with confidence under 50%. Although different thresholds lead to higher final F1 scores, we report results with confidence threshold of 50% for a fair comparison with previous research.

4.2 Baselines

We compare LIME with both entailment-based classification (Phase 1 without self-training) and previous research on weakly-supervised text classification. We also include BERT trained with supervision from original labels as a realistic upper bound for weakly-supervised classification.

WestClass (Meng et al., 2018) generates pseudo-documents for each class label. **ConWea** (Mekala and Shang, 2020) utilizes a pre-trained language model to discern keywords that carry different meanings under different contexts. **LotClass** (Meng et al., 2020b) is a framework for text classification using only label names. **LotClass** mines a pre-trained BERT model for seed words that are most likely to replace each class name. **X-Class** (Wang et al., 2021) is a state-of-the-art weakly-supervised classification system that collects seed words within the test documents instead of external sources. Documents are then grouped with a Gaussian Mixture Model before pseudo-labels are assigned.

4.3 Datasets

We run LIME on standard benchmarks in weakly-supervised classification: **20News** (Lang, 1995), **AGNews** (Zhang et al., 2015), **Yelp reviews** (Zhang et al., 2015), and **DBpedia** (Zhang et al., 2015). Detailed descriptions of each dataset, along with specific prompts used, are recorded in Table 2. We notably omit NYT datasets used in Meng et al. (2020a) and Wang et al. (2021), because only pre-processed (all lower-cased, pre-tokenized with a specific tokenizer) versions of the data were available. It is not possible to meaningfully evaluate the pseudo-labeling scheme in LIME if test samples are tokenized by a tokenizer different from that coupled with our entailment model.

5 Results

5.1 Classification performance

Final classification results are recorded in Table 3. LIME outperforms all baselines in terms of micro- and macro-F1 scores, even approaching the supervised baseline in the **Yelp** dataset.

We also find that training a new classifier with pseudo-labels (Phase 2 of LIME) does not amplify or propagate errors in incorrect pseudo-labels. The final classifier consistently scores roughly 10 points

[2]https://huggingface.co/facebook/bart-large-mnli

Model	20News	AGNews	Yelp	DBpedia
Supervised	96.45 / 96.42	93.99 / 93.99	95.70 / 95.70	98.96 / 98.96
Entailment classifier	67.95 / 67.50	79.94 / 79.99	94.79 / 94.79	80.14 / 79.27
WeSTClass	71.28 / 69.90	82.30 / 82.10	81.60 / 81.6	81.42 / 81.19
ConWea	75.73 / 73.26	74.60 / 74.20	71.40 / 71.20	N/A
LOTClass	73.78 / 72.53	86.89 / 86.82	87.75 / 87.68	86.66 / 85.98
X-Class	78.62 / 77.76	85.74 / 85.66	90.00 / 90.00	91.32 / 91.17
LIME	**79.74 / 79.56**	**87.21 / 87.16**	**95.22 / 95.22**	**92.19 / 92.20**

Table 3: Experiment results on 4 classification benchmarks. All reported scores in the form *micro-F1 / macro-F1*. Baselines are quoted from (Wang et al., 2021).

higher in F1 scores compared to the entailment classifier. Our results confirm findings from previous research that employ self-training to improve classification robustness in low-resource regimes (Mukherjee and Awadallah, 2020; Gowal et al., 2021).

5.2 Effect of label confidence thresholds

Figure 1 plots the spread of pseudo-label confidence produced in Phase 1 of LIME. We confirm that higher average confidence from the entailment classifier in Phase 1 robustly translates to higher classification accuracy for both the entailment classifier and the self-trained classifier in Phase 2.

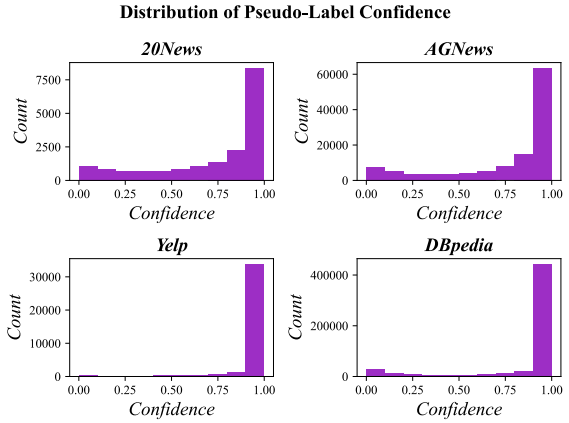

Figure 1: Histogram of pseudo-label confidence. More confident pseudo-labels result in more accurate classification self-training.

Another notable finding is that naively utilizing only high-confidence labels does not always guarantee a more accurate classifier. A trade-off exists between filtering out low-confidence labels and retaining a sizable training corpus. We find that confidence cut-off from 50% to 70% strikes a good balance between the two obligations (Figure 2).

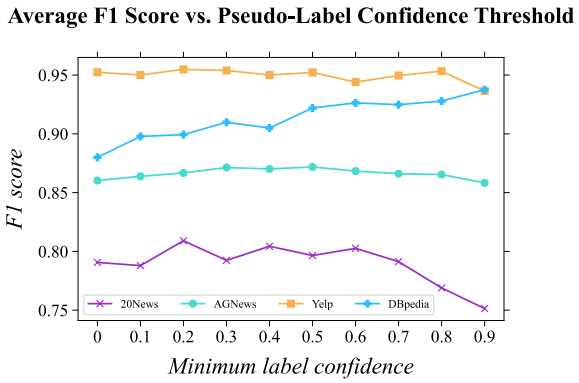

Figure 2: Effect of varying confidence thresholds on self-training F1 scores.

6 Conclusions

LIME proposes a streamlined pseudo-labeling method for weakly-supervised text classification. The framework combines flexibility of entailment-based classification with robustness of self-training. The resulting text classifier outperforms previous state-of-the-art in weakly-supervised classification. We also investigate the effect of pseudo-label confidence thresholds on self-trained classifier performance. Entailment model confidence accurately reflects label accuracy, but size of the pseudo-training set is also important for robust classification.

We identify several avenues for future research. For a fair comparison with previous research, we did not modify the self-training step with more advanced neural classifier architectures or confidence-aware self-training schemes (Mukherjee and Awadallah, 2020). Other auxiliary tasks, such as question-answering (McCann et al., 2018) or next sentence prediction (Ma et al., 2021) can also extend the LIME framework as alternate pseudo-classifiers.

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yaroslav Fyodorov, Yoad Winter, and Nissim Francez. 2000. A natural logic inference system.

Sven Gowal, Po-Sen Huang, Aaron van den Oord, Timothy Mann, and Pushmeet Kohli. 2021. Self-supervised adversarial robustness for the low-label, high-data regime. In *International Conference on Learning Representations*.

Ken Lang. 1995. Newsweeder: Learning to filter netnews. In Armand Prieditis and Stuart Russell, editors, *Machine Learning Proceedings 1995*, pages 331–339. Morgan Kaufmann, San Francisco (CA).

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Tingting Ma, Jin-Ge Yao, Chin-Yew Lin, and Tiejun Zhao. 2021. Issues with entailment-based zero-shot text classification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 786–796, Online. Association for Computational Linguistics.

Bill MacCartney and Christopher D. Manning. 2009. An extended model of natural logic. In *Proceedings of the Eight International Conference on Computational Semantics*, pages 140–156, Tilburg, The Netherlands. Association for Computational Linguistics.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering.

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.

Yu Meng, Jiaxin Huang, Guangyuan Wang, Zihan Wang, Chao Zhang, Yu Zhang, and Jiawei Han. 2020a. *Discriminative Topic Mining via Category-Name Guided Text Embedding*, page 2121–2132. Association for Computing Machinery, New York, NY, USA.

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, Yunyi Zhang, Jiaxin Huang, Chenyan Xiong, Heng Ji, Chao Zhang, and Jiawei Han. 2020b. 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.

Subhabrata Mukherjee and Ahmed Awadallah. 2020. Uncertainty-aware self-training for few-shot text classification. In *Advances in Neural Information Processing Systems*, volume 33, pages 21199–21212. Curran Associates, Inc.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, 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*, pages 3043–3053, Online. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3914–3923, Hong Kong, China. Association for Computational Linguistics.

Wenpeng Yin, Nazneen Fatema Rajani, Dragomir Radev, Richard Socher, and Caiming Xiong. 2020. Universal natural language processing with limited annotations: Try few-shot textual entailment as a start. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8229–8239, Online. Association for Computational Linguistics.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.