De-Noising Document Classification Benchmarks via Prompt-based Rank Pruning: A Case Study

Matti Wiegmann[1], Benno Stein[1], and Martin Potthast[2]

[1] Bauhaus-Universität Weimar, Germany
[2] University of Kassel, hessian.AI, and ScaDS.AI, Germany

Abstract Model selection is based on effectiveness experiments, which in turn are based on benchmark datasets. Benchmarks for "complex" classification tasks, such as tasks with a high subjectivity, are prone to label noise in their (manual) annotations. For such tasks, experiments on a given benchmark may therefore not reflect the actual effectiveness of a model. To address this issue, we propose a three-step de-noising strategy: Given labeled documents from a complex classification task, use large language models to estimate "how strong the signal within a document is in the direction of its class label", rank all documents according to their estimated signal strengths, and omit documents below a certain threshold. We evaluate this strategy in a case study on the assignment of trigger warnings to long fan fiction texts. Our analysis reveals that the documents retained in the benchmark contain a higher proportion of reliable labels, and that model effectiveness assessments are more meaningful and models become easier to distinguish.[1]

1 Introduction

There are text classification tasks for which providing a sufficient amount of labeled data is difficult. The difficulty may be due to the subjectivity of the task (Is this text a product *description* or a product *advertisement*?), a high number of classes (Which of the 188 cognitive biases occur in this text?), a missing dichotomy since only one class can be characterized (Does this text has an enticing writing style?), the need for expert knowledge (Is argument A more convincing than argument B?), or a combination of these characteristics. For such tasks, LLMs have shown great performance, even in zero-shot settings.

But, just as powerful as LLMs are in this respect, they are obviously not a panacea: Time, cost, and latency are among their main limiting factors, especially for classification tasks that require ad hoc decisions and high throughput. Consider, for example, the generation of a search engine result page (SERP) on which documents containing product advertising, undesirable prejudices, or sarcasm are to be filtered out. The practical and efficient approaches, instead, fine-tune neural networks based on dense document representations , such as BERT or RoBERTa [8]. Their limiting factor, however, is the knowledge acquisition bottleneck, i.e. the lack or the quality of labeled data. This lack of labeled

[1] Code and Data: https://github.com/webis-de/CLEF-24

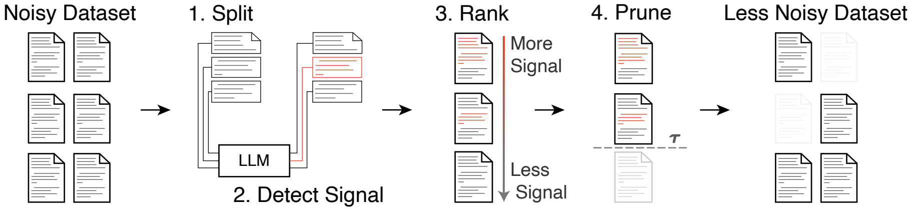

Figure 1. Overview of the proposed method of pruning documents with a label depending on how strong the signal for this label is according to an LLM classifier.

data is often countered by collecting data from weakly-supervised sources. One example of this is the extraction of trigger warnings from online blogs, where authors signal if their work contains harmful content.

However, weakly-supervised data acquisition leads to noisy data due to errors or inconsistencies in the distant knowledge source. The use of noisy data to benchmark classification models (which is the focus of this paper) is problematic: model performances may be underestimated, model differences may be smaller or vanish, or, in the worst case, leaderboard rankings change. Or the other way around: reducing label noise in benchmark data increases model scores and may increase the performance difference between models, which makes it easier to assess which model is actually better and by how much.

This is where our contribution comes in: The paper in hand proposes prompt-based text classification to reduce label noise, especially false positives, in difficult document classification benchmarks (i.e. test datasets) (cf. Figure 1). We use the LLMs to detect how much signal is present in each document to justify the label assigned to it, and we remove the documents with the weakest signal (Section 3). We evaluate our method using three common models (XGBoost, RoBERTa, Longformer) on a multi-label trigger detection dataset [17] (as used in a joint task on CLEF 2023 [16]), which provides some organic information about label reliability (Section 4).

Our results (Section 5) show that our method increases the ratio of noisy to reliable documents in the benchmark from 1:1 up to 1:6, that models tested on de-noised data score up to 0.15 F_1 higher than when tested on "noisy" documents, and that models may scores the same on noisy data but significantly different on de-noised dataset.

2 Related Work

Although current (pre-trained) deep learning models are somewhat robust to label noise given sufficient training data [13,19], reducing label noise is still essential when training non-neural models [9,3] or with limited training data. Most related work focuses on training data de-noising neural classifiers [18,6], especially with semi-supervised methods like adapting the loss function [11,14], by over-parameterization [7], or by rank pruning [10] via predicted probabilities. Some related works also use weak supervision methods to estimate label relia-

Table 1. Number and length of eligible source and sampled evaluation documents.

Warning	**Source Data**		**Sample used in this Work**			**Length**	
	Unknown	Reliable	Unknown	Flipped	Reliable	Mean	Std
Death	124,958	1,579	600	200	200	3,351	2,717
Violence	119,684	1,736	600	200	200	4,021	2,853
Homophobia	22,688	558	600	200	200	4,125	2,809
Self-harm	23,029	1,343	600	200	200	3,478	2,688

bility [5,12] from (multiple) external sources. For our work, we adapt the rank pruning idea but use an external source (an LLM) instead of a semi-supervised signal. However, the most notable difference of our work is that we do not focus on de-noising the training data to improve the model but the test data to improve the benchmark reliability, which is why we study organic noise instead of only injecting synthetic noise like the related work (e.g., on TREC question-type and AG-News datasets [4]).

3 Finding and Pruning Noisy Documents

Our label de-noising procedure assumes the following: First, the input dataset contains a set of documents, and each document has one or more labels from a finite set. Second, each reliable document with a true positive label contains a signal above a confidence threshold τ (i.e., a piece of set) that justifies the label. Third, there are a number of noisy documents that have been assigned a positive label where the signal with respect to that label is weaker than τ. Our pruning strategy, illustrated in Figure 1, attempts to find and remove documents that are noisy with respect to a particular label by determining the signal strength of that label.

To do this, we rank all documents independently for each label according to the strength of the signal of this label and then determine τ as a threshold. The de-noising scheme consists of four steps for each label: (1) Splitting of documents into smaller chunks, i.e. several consecutive sentences, where the chunk size is a hyperparameter. (2) Determine whether a chunk carries a signal for the label using a prompt-based binary classification, where LLM and prompt are hyperparameters that depend on the task and the label. (3) Ranking of the documents on the basis of the absolute number of signals, i.e. the positively classified chunks. (4) Pruning of the documents with the lowest rank up to a rank or signal strength threshold τ.

4 Experimental Evaluation

We evaluate LLM-based benchmark de-noising on a multi-label classification task and evaluate the noise ratio and model effectiveness at different τ.

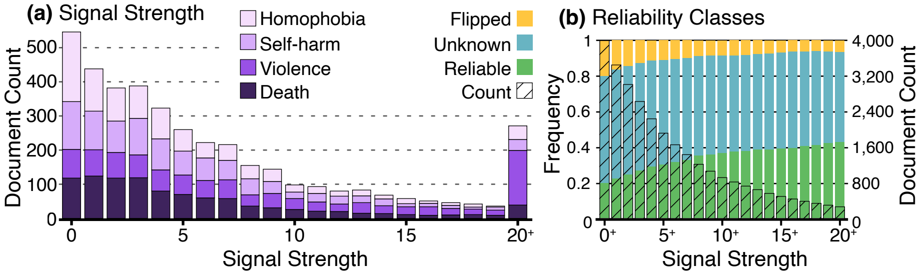

Figure 2. **(a)** Signal strength distribution: Number of documents with a certain amount of positively classified 5-sentence chunks by label. **(b)** Number of documents and distribution of document reliability in the pruned corpus at different thresholds.

Dataset We use evaluation data from the Webis Trigger Warning Corpus (WTWC) [17], which was used in the 2023 shared task on trigger detection [16]. The WTWC is well suited as it contains organic false positive and negative labels that emerge from human authors (sensitive human authors assign warnings for weak signals) and from weakly supervised labeling (which assigns warnings for loosely related or implied concepts). The dataset also contains additional reliability information in the "author notes" prepended to some chapters.

We sample 4,000 WTWC documents balanced across 4 warning labels *Death*, *Violence* (the two most common warnings, excluding Pornography as outlier), *Homophobia* and *Self-harm* (the two closest to median frequency with sufficient *Reliable* documents) as our evaluation dataset (cf. Table 1), which is large enough to test our method. For each label, we first sample 200 *Reliable* documents where the author note mentions either `tw, cw, trigger(s), content warning` within 20 tokens of a warning term (e.g. `homophobia`). Then, we sample 800 non-*Reliable* documents and create a subset of 200 known falsely labeled data by *Flipping* the documents' label to a different one. The reliability of the remaining 600 documents was marked as *Unknown*. We adopted all other sampling criteria from the shared task [16] (English documents; 50-10,000 words; no duplicates).

4.1 De-Noising Implementation

We apply our de-noising technique (Section 3) using 5 consecutive sentences without overlap as chunks and *Mixtral-8x7B-v0.1* from Huggingface as LLM. We use a binary classification prompt aligned with Mistral's prompting guide:

```
You are a text classification model. You determine if a given text contains
death, graphic display of death, murder, or dying characters. If the given
text contains intense, explicit, and graphic death, you answer: Yes. If the
text contains mild or implicit death or no death at all, you answer: No.
```

We classify by predicting the next-token probabilities and comparing the logits of the `Yes` and `No` tokens. We rank and prune the documents by the absolute

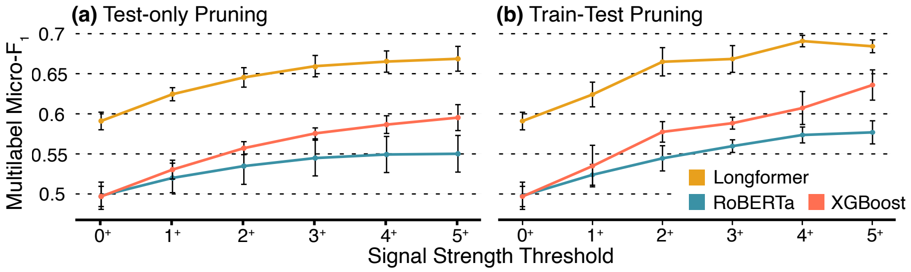

Figure 3. Model F_1 with confidence intervals of three classification models at different pruning thresholds when (a) only test data and (b) training and test data are pruned.

number of positively (`Yes` > `No`) classified chunks per document, i.e. at a τ of 5^+ all documents with less than 5 positive chunks will be pruned.

4.2 Experiments and Evaluation

To evaluate our hypotheses we conduct three experiments across three baseline classification models. First, we prune the complete dataset (with τ from 0^+ to 20^+) and observe the ratio of reliability classes. Second, we split the data 80:20 into training and test and only prune the test dataset (with τ from 0^+ to 5^+)[2] while training the models on the complete training data. Third, we prune the complete dataset (with τ from 0^+ to 5^+) before the train-test split and also train the models with pruned data. Decreasing scores in this last experiment would indicate that our method also removes (many) difficult cases, leading to both, poor models and a poor benchmark.

We train three models for multi-label classification: a fine-tuned `FacebookAI/roberta-base` and `allenai/longformer-base-4096` [1] and a feature-based `XGBoost` [2] classifier (the baseline of the shared task [16]) with the top 10,000 tf · idf word 1–3-gram features selected via χ^2. The RoBERTa input was truncated to 512 tokens and the Longformer input to 4,096 tokens. We report the micro-averaged multi-label F_1 via a 5-fold Monte Carlo cross-validation and the 95% t-estimated confidence intervals. Our code repository lists training parameters and our ablation study.

5 Results and Discussion

Our first assumption is that our method removes noise from the dataset if, with increasing τ, the proportion of *Reliable* documents increases and of *Flipped* documents decreases. Figure 2(b) shows that the proportion of *Reliable* documents increases from 0.2 to 0.41 and decreases for *Flipped* documents from 0.2 to 0.05. Note that the proportion changes are strongest for smaller τ.

Our second assumption is that de-noising improves the benchmark when the models' test scores increase with increased de-noising (for train-test and test-only

[2] At $\tau = 5^+$, half the dataset has been pruned.

pruning) and when the relative difference between models' test scores changes. Figure 3(a) shows that the F_1 of all models increases by 0.05–0.1 with $\tau = 5^+$ when pruning only the test data. The effect is strongest for XGBoost and weakest for RoBERTa (where the input documents are strongly truncated). Figure 3(a) also shows that XGBoost and RoBERTa score evenly without pruning but XGBoost improves more strongly and is significantly more effective with $\tau = 5^+$. This shows that de-noising can reveal model differences that are otherwise hidden by the noise. Figure 3(b) shows that the F_1 of all models increases when pruning all data and more strongly than when only pruning the test data.

6 Conclusion

In this paper, we investigate using rank-based pruning based on an LLMs classification signal to de-noise a document-level trigger warning classification dataset. We show that our de-noising strategy doubles the relative number of reliably labeled documents and halves the noisily labeled ones. We further show that our de-noising strategy increases the model scores and the differences between models, hence we assume that the de-noised dataset is more suited as a benchmark.

References

1. Beltagy, I., Peters, M., Cohan, A.: Longformer: The Long-Document Transformer. CoRR (2020)
2. Chen, T., Guestrin, C.: XGBoost: A Scalable Tree Boosting System. 22nd SIGKDD (2016)
3. Frénay, B., Verleysen, M.: Classification in the Presence of Label Noise: A Survey. IEEE Transactions on Neural Networks and Learning Systems **25**, 845–869 (2014)
4. Garg, S., Ramakrishnan, G., Thumbe, V.: Towards Robustness to Label Noise in Text Classification via Noise Modeling. 30th ACM CIKM (2021)
5. J. Ratner, A., de Sa, C., Wu, S., Selsam, D., Ré, C.: Data Programming: Creating Large Training Sets, Quickly. Advances in neural information processing systems **29**, 3567–3575 (2016)
6. Liu, S., Niles-Weed, J., Razavian, N., Fernandez-Granda, C.: Early-Learning Regularization Prevents Memorization of Noisy Labels. 33rd NeurIPS (2020)
7. Liu, S., Zhu, Z., Qu, Q., You, C.: Robust Training under Label Noise by Over-parameterization. ArXiv **abs/2202.14026** (2022)
8. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized BERT pretraining approach. CoRR (2019)
9. Natarajan, N., Dhillon, I., Ravikumar, P., Tewari, A.: Learning with Noisy Labels. In: Neural Information Processing Systems (2013)
10. Northcutt, C.G., Wu, T., Chuang, I.L.: Learning with Confident Examples: Rank Pruning for Robust Classification with Noisy Labels. ArXiv (2017)
11. Patrini, G., Rozza, A., Menon, A.K., Nock, R., Qu, L.: Making Deep Neural Networks Robust to Label Noise: A Loss Correction Approach. 2017 IEEE CVPR (2016)
12. Ren, W., Li, Y., Su, H., Kartchner, D., Mitchell, C.S., Zhang, C.: Denoising Multi-Source Weak Supervision for Neural Text Classification. ArXiv (2020)
13. Rolnick, D., Veit, A., Belongie, S.J., Shavit, N.: Deep Learning is Robust to Massive Label Noise. ArXiv **abs/1705.10694** (2017)
14. Sanchez, E.A., Ortego, D., Albert, P., O?Connor, N.E., McGuinness, K.: Unsupervised label noise modeling and loss correction. ArXiv **abs/1904.11238** (2019)
15. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E.H., Le, Q.V., Zhou, D.: Chain-of-thought prompting elicits reasoning in large language models (2022)
16. Wiegmann, M., Wolska, M., Potthast, M., Stein, B.: Overview of the Trigger Detection Task at PAN 2023. Working Notes of CLEF. CEUR-WS, vol. 3497, pp. 2523–2536 (Sep 2023)
17. Wiegmann, M., Wolska, M., Schröder, C., Borchardt, O., Stein, B., Potthast, M.: Trigger Warning Assignment as a Multi-Label Document Classification Problem. 61th ACL (2023)
18. Zhang, Z., Zhang, H., Arik, S.Ö., Lee, H., Pfister, T.: Distilling Effective Supervision From Severe Label Noise. 2020 IEEE/CVF CVPR (2019)
19. Zhu, D., Hedderich, M.A., Zhai, F., Adelani, D.I., Klakow, D.: Is BERT Robust to Label Noise? A Study on Learning with Noisy Labels in Text Classification. ArXiv **abs/2204.09371** (2022)