Report on arXiv Scientific Text Classification

Jordy Van Landeghem[1] [1]KU Leuven

1. Summary

This report presents the results of the arXiv Scientific Text Classification task. The goal of this task is to classify scientific papers into one or more categories based on the paper's abstract. The dataset consists of +2.5M papers, which are split into a category-stratified training set of 2M papers and a validation set of 500K papers.

First, I focused on establishing a discriminative encoder-based baseline for multi-label classification using BERT [2]. I then experimented with different encoders (DeBERTa [3] (known to be a more powerful encoder), SciBERT [7] (in-domain pretraining)), batch sizes (6→64). learning rates (1e-4→2e-5), and loss functions (binary cross-entropy (CE) vs. two-way multi-label loss [5]) to improve the model's performance.

Out of interest, I also explored alternate approaches such as few-shot classification with a one-vs-rest strategy (SetFit [10]) and generative models (Llama2 [9]) instruction-tuned on human-readable labels. The report concludes with a discussion of future work. In total, I invested <20 hours in this task and mainly focused on prototyping and experimenting with different approaches at multi-label text classification.

The key observations of this report are:

I. Multi-label classification with a large dataset is challenging due to label imbalance and label noise. More methods exist to deal with this for discriminative models, but few-shot learning and generative models are interesting alternatives.

II. Certainly, the choice of pre-trained model and loss function is crucial for the performance of the model. The two-way multi-label loss function outperformed the baseline binary cross-entropy loss function. (TBC)

2. Exploratory Data Analysis

The notebook 'EDA.ipynb' contains documentation and visualization of observations made on the arXiv dataset.

The papers are classified into 149 unique categories with a frequency larger than 10 and with a human-readable label available in the arXiv category taxonomy, with a mean of 1.6958 categories per paper. The distribution of the number of papers per category is highly skewed. As part of the preprocessing, all non-unique titles were removed and outlier abstracts (length < 100 or > 1000) were removed. Using 'langdetect' I found that the abstracts were predominantly in English, which informs the choice of pretrained models. The dataset is highly imbalanced, with a long-tailed distribution of the number of papers per category. To enable stratified validation and test splits, I bucketed all multi-label categories together that had less than 10 papers.

The preprocessed dataset is available at https://huggingface.co/datasets/jordyvl/arXiv_dataset_prep. The dataset contains the following fields:

- *abstract*: without any preprocessing such as lemmatization, risky with many technical terms, could reduce overfitting when done right
- *primary*: the primary category of the paper; could be used for multi-class classification
- *categories*: the categories of the paper as a list of labels, *e.g.* [cs.AI, eess.AS]; multi-label classification
- *strlabel*: human-readable labels for the categories joined with ';', *e.g.* [cs.AI, eess.AS] → "Artificial Intelligence;Audio and Speech Processing"

3. Experiments

3.1. Methods

Some motivation is due for the choice of alternate approaches than encoder-based discriminative models. The discriminative models are known to be powerful and are the current baselines of choice for multi-label classification. However, they are also known to be data-hungry, which is hard in the case of an imbalanced long-tailed label distribution as here, and require substantial fine-tuning to achieve good performance.

A generative pre-trained LLM such as Llama2 [9] is known to be more flexible and can be instruction tuned toward any task described in natural language with potentially fewer instances, while rendering a semantic understanding of the labels and even the ability to generate new labels. Given the higher computational cost involved with LLMs, I subsampled the dataset to 10% of the original size to keep the computation time reasonable.

SetFit [10] is a few-shot classification framework for fine-tuning Sentence Transformers. It is substantially more data-efficient and can be trained with different multi-label strategies (one-vs-rest), which might be more robust to label noise and label imbalance. I opted for this setup, as in industry settings, supervised learning is often not feasible due to the lack of a large set of labeled data, and few-shot learning is a promising alternative. I created a new Sentence Transformer from SciBERT and used this to fine-tune the SetFit model.

I also experimented with different loss functions, such as the two-way multi-label loss [5], which was reported to outperform other known loss functions (focal loss, assymmetric loss, ...) on the multi-label classfication task. It is inspired by the properties of softmax-based CE (as opposed to sigmoid binary CE) and contrastive learning over samples and labels.

3.2. Evaluation

The performance of the discriminative models is evaluated using standard metrics: *precision, recall, micro-averaged F1 score, and hamming loss*. The generative models are evaluated using my own extended metric, *Average Normalized Levenshtein Similarity* (ANLS) [11], that evaluates the quality of the generated labels and the similarity to the human-readable labels, agnostic of the order of the labels and permissive to low edit distance on individual generated labels.

For example, the ANLS of the generated labels "Audio and Speech Processing;Artificial Intelligence" and the human-readable labels "Artificial Intelligence;Audio and Speech Processing" is 1.0, as well as for "Audio Processing;Artificial Intelligence" given an NLS threshold τ of 0.5. The threshold can be adapted to the task's needs, *e.g.* to enforce a minimum similarity to the human-readable labels, such that even incomplete ones could be reconstructed by surface form or semantic similarity (*e.g.* Audio Processing is closest to Audio and Speech Processing of all other labels).

4. Results

All results are reported on a test set subsampled from the validation set to keep computation time reasonable. The best attained results are reported in Table 1. All experiment runs and results are available at `https://wandb.ai/jordy-vlan/scientific-text-classification`.

5. Future Work

Of course, all the following depends on the needs of the task and the available resources.

- ☐ **Hyperparameter tuning**: extend MultiLabelTrainer with Optuna or Ray Tune or Wandb sweeps
- ☐ **Feature fusion** from arXiv metadata (*e.g.* authors (co-citation network), date of submission)
- ☐ **Ensembling** of different pre-trained models (*e.g.* BERT, DeBERTa, SciBERT, SetFit, Llama2), potentially weighted by normalized validation scores

Table 1. Results of the different models (trained with different number of steps based on loss curve validation) on the test set. The hamming loss is minimized, while other metrics are maximized.

Model	Accuracy	Precision	Recall	F1	Hamming(↓)	ANLS
BERT(10K)	0.993	0.776	0.461	0.578	0.007	
BERT	0.993	0.776	0.461	0.578	0.007	
DeBERTa (50K)	0.994	0.763	0.555	0.643	0.006	
SciBERT	0.994	0.797	0.525	0.633	0.006	
SciBERT (2-way)	0.994	0.791	0.577	0.667	0.006	
SciBERT(100K)*	**0.995**	**0.779**	**0.627**	**0.695**	**0.005**	
SetFit (ovr@20)						
SetFit (ovr-diff@20)						
Llama2						0.607

  - Combine predictive models with **different output spaces** (*e.g.* multi-label, multi-class) to enforce consistency on the primary category and average label cardinality

☐ Continue with **encoder-decoder models** for multi-label classification (*e.g.* SGM [**?**] T5Enc [4])

  - Comment: overkill for single task finetuning, could be useful when combining multiple tasks (summarization, translation, classification, . . .)

☐ More advanced **prompt design** with function calling for postprocessing the output of LLMs (*e.g.* JSON structure for unique classes)

☐ Further explore the SetFit framework for few-shot multi-label classification, albeit it is admittedly not well-designed for this task and might require substantial adaptation

☐ Investigate graph neural networks for better use of the **label hierarchy** (*e.g.* in the medical domain [1])

☐ Explore strategies for dealing with **label noise** (*e.g.* label smoothing, CleanLab [6, 8])

References

[1] Shengqiang Chi, Yuqing Wang, Ying Zhang, Weiwei Zhu, and Jingsong Li. Graph neural network based multi-label hierarchical classification for disease predictions in general practice. *Studies in Health Technology and Informatics*, 310:725–729, 2024. 3

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*, pages 4171–4186, 2018. 1

[3] Pengcheng He, Xiaodong Liu, Jianfeng Wang, Weizhu Li, Yelong Liu, Xiuying Zhao, Xinyan Xiao, Jiawei Liu, and Yeyun Lyu. Deberta: Decoding-enhanced bert with disentangled attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 156–168, 2020. 1

[4] Yova Kementchedjhieva and Ilias Chalkidis. An exploration of encoder-decoder approaches to multi-label classification for legal and biomedical text. *arXiv preprint arXiv:2305.05627*, 2023. 3

[5] Takumi Kobayashi. Two-way multi-label loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7476–7485, 2023. 1, 2

[6] Himanshu Kumar, Naresh Manwani, and PS Sastry. Robust learning of multi-label classifiers under label noise. In *Proceedings of the 7th ACM IKDD CoDS and 25th COMAD*, pages 90–97. 2020. 3

[7] Himanshu Maheshwari, Bhavyajeet Singh, and Vasudeva Varma. Scibert sentence representation for citation context classification. In *Proceedings of the Second Workshop on Scholarly Document Processing*, pages 130–133, 2021. 1

[8] Diane Oyen, Michal Kucer, Nicolas Hengartner, and Har Simrat Singh. Robustness to label noise depends on the shape of the noise distribution. *Advances in Neural Information Processing Systems*, 35:35645–35656, 2022. 3

[9] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 1, 2

[10] Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. Efficient few-shot learning without prompts. *arXiv preprint arXiv:2209.11055*, 2022. 1, 2

[11] Jordy Van Landeghem, Rubèn Tito, Łukasz Borchmann, Michał Pietruszka, Pawel Joziak, Rafal Powalski, Dawid Jurkiewicz, Mickaël Coustaty, Bertrand Anckaert, Ernest Valveny, Matthew Blaschko, Marie-Francine Moens, and Tomasz Stanisławek. Document Understanding Dataset and Evaluation (DUDE). In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19528–19540, 2023. 2