Team Elsevier Health Data Science at BioCreative VII LitCovid Track: Fine-Tuning Transformer Models for COVID-19 Literature Annotation

A Multi-Class Classification Problem

Tabitha Sugumar, McCullen Sandora, Marvin Thielk, Sharvari Jadhav, Sameer Chivukula
Affiliation: Elsevier

***Abstract*— The vast and growing body of Covid-19 literature poses a challenge for parsing, interpretation, and curation. In order to address that, the BioCreative Challenge called for approaches to automate topic classification, based on the topics and current labelled data available in the LitCovid database. Our approach involved fine-tuning pre-trained transformer models with a classification head to label Covid-19 literature according to the LitCovid database's topic framework. The models we experimented with included Distilbert, Longformer, and Scibert. We also experimented with fine-tuning on different features of the training data – the article abstracts, article titles, and all the information together. The best performing model was a version of Scibert pre-trained on a corpus including Covid-19 papers. This model was then fine-tuned on the title, abstracts, keywords, and journal name concatenated together, with weighting during the training process to account for class imbalances.**

Keywords—multi-class classification; fine-tuning; language models; transformers

I. Introduction

The LitCovid database (4) hosts Covid-19 related papers from PubMed. The database's topic labelling scheme involves eight categories: Case Report, Diagnosis, Epidemic Forecasting, Mechanism, Prevention, Transmission, and Treatment, and each article is labelled with one or more topics. These topics have been useful in information retrieval and are part of many applications of the database, however manually labelling each article has impeded the process of updating the database.

This BioCreative challenge (3) provided a dataset of labelled articles, consisting of their titles, abstracts, keywords, authors, journal name and the manually annotated topics. Structured as a multiclass classification problem, the standard approach of fine-tuning a pre-trained, out-of-the-box transformer model with a classification head was taken.

II. Methods

The pre-trained models were accessed and then fine-tuned using the Huggingface library. The standard classification head provided by Huggingface was used, and the model was trained with a sigmoid loss function, standard for multiclassification. The following models were tested:

- Distilbert: Huggingface's model that emulates BERT at a much smaller size, speeding up frequency of training (5).
- Longformer: Transformer model that allows for many tokens while scaling training time linearly by manipulating the attention mechanism. This makes processing of larger text sizes possible, but without the complete attention matrix performance is often negatively impacted. When tested with this dataset the maximum number of tokens was set to 1024, as opposed to 512 tokens used for the other models (2).
- Covid-Scibert: The original Scibert was trained on a corpus of scientific journal extracts (Wolf 2019). This model extends the original Scibert with further pretraining on a corpus of journals related to Covid-19 (6).

In addition, combinations of title, abstract, keywords, and journal name were used as inputs. The combinations were created by concatenating the strings together. A single test was done using the title and abstract separated with Bert's SEP token.

Additional experiments were done to account for the imbalance of classes in the problem. Approaches included weighting during training based on proportions of classes in the training data.

III. Results

The results in Table 1 show the performance of a selection of experiments on the provided development set. Comparing the performance of the Distilbert model fine-tuned on the abstracts, with and without weights shows that including weighting improves the F1 score by 0.0043. A closer examination of precision and recall indicates that a weighted training approach causes precision to drop by 0.01 while recall increases by 0.02, therefore the significance of the type I and type II errors are something to consider.

TABLE I. MODEL PERFORMANCE COMPARISON

Features	Model	Weighting	Instance F1	Micro F1
Title, Abstract, Journal Name, Keywords	Covid-Scibert	Weights	0.9289	0.9027
Abstract	Longformer	Weights	0.9122	0.8805
Abstract	Distilbert	Weights	0.9169	0.8894
Title	Distilbert	Weights	0.8308	0.7951
Title, SEP, Abstract	Distilbert	Weights	0.8481	0.8145
Abstract	Distilbert	None	0.9126	0.8898

Using Longformer dropped the F1 score by 0.0047 in comparison to Distilbert, indicating that the additional information from increasing the number of tokens possible (from a maximum of 512 to 1024) does not have a significant enough impact in this case to make up for the simplification of the information captured by the attention mechanism.

The best performing model tested used a version of Scibert, pre-trained on a corpus of journal articles related to Covid-19 on top of the original Scibert model (1), pre-trained on a corpus of papers from Semantic Scholars. This model was then fine-tuned on the title, abstract, journal name and keywords concatenated together, with weighting based on the proportion of each class in the training data. On the official test set the average instance F1 score for this model was 0.9307, the average instance precision was 0.9244, and the average instance recall was 0.9371.

In the context of the challenge, our model achieves an instance based F1 score 0.0053 greater than the third quartile of team submissions, with a higher precision and lower recall. The label based F1 score of our model is 0.0009 less than the third quartile of submissions, suggesting that better handling of class imbalances may be worth pursuing for our approach. Table II summarizes the comparative performance of our model with that of the challenge's baseline and team submissions.

TABLE II. Fine-tuned Covid-Scibert Performance

	Label-based Micro Average			Instance-based		
	Prec.	**Recall**	**F1**	**Prec.**	**Recall**	**F1**
Fine-tuned Covid-Scibert	0.8979	**0.9170**	0.9074	0.9244	**0.9371**	**0.9307**
Team Submissions						
Q3	**0.9251**	0.8964	**0.9083**	**0.9353**	0.9192	0.9254
Mean	0.8967	0.8624	0.8778	0.8985	0.8887	0.8931
Median	0.9108	0.8843	0.8925	0.9188	0.9097	0.9132
Std	0.0541	0.0482	0.0429	0.0521	0.0451	0.0460
Baseline (ML-Net)	0.8756	0.8142	0.8437	0.8849	0.8514	0.8678

IV. CONCLUSION

By fine-tuning out of the box, pre-trained language models on the data provided, we were able to attain reasonable results for labelling Covid-19 papers based on the LitCovid database's topic framework. This approach is a standard strategy for multi-classification problems, however a few points to note did come up in our experiments.

One challenge in this dataset is the imbalanced class distribution – some topics come up far more often than others. Weighting classes based on their proportion in the dataset during training was the most effective approach we attempted. Given that this improved performance, it stands to reason that a more nuanced approaches to handle class imbalance could be worth trying.

Finally, the most successful pretrained model was pre-trained on a corpus including journal articles discussing Covid-19. This is unsurprising as the problem at hand involves classifying Covid-19 literature. Given that the literature related to Covid-19 is increasing at a rapid rate, and is in fact the reason this classification problem needs to be solved at all – if this approach to classification is taken going forward it may make sense to further pre-train the language model on a larger corpus of Covid-19 journal articles, before fine-tuning for the multi-classification task.

REFERENCES

1. Beltagy,I., Lo,K. and Cohan,A. (2019) SciBERT: A Pretrained Language Model for Scientific Text. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3615-3620.

2. Beltagy,I., Peters,M.E. and Cohan,A. (2020) Longformer: The Long-Document Transformer. *ArXiv,* 2004.05150v2**.**

3. Chen,Q., Allot,A., Leaman,R., Doğan,R.I. and Lu,Z. (2021) Overview of the BioCreative VII LitCovid Track: multi-label topic classification for COVID-19 literature annotation. *Proceedings of the seventh BioCreative challenge evaluation workshop.*

4. Chen,Q., Allot,A. and Lu,Z. (2020) Keep up with the Latest Coronavirus Research. *Nature,* 193.

5. Sanh,V., Debut,L., Chaumond,J. and Wolf,T. (2020) DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv,* 1910.01108.

6. Thakur,T. (2021) Hugging Face, COVID-SciBERT. Accessed Oct 31, 2021. https://huggingface.co/lordtt13/COVID-SciBERT.