Team TCSR at BioCreative VII LitCovid Track: Automated topic prediction of LitCovid using BioBERT

Saipradeep VG, Naveen Sivadasan, Aditya R Rao, Thomas Joseph
TCS Research, Hyderabad, INDIA

Abstract— **Background: The BioCreative VII- Track 5 challenge aims to increase the accuracy of automated topic prediction in the massively growing COVID-19 literature with the aim of helping research community to find effective diagnostics, drugs and vaccines for COVID-19. The challenge requires the participants to perform multi-label assignment task where each of the 33.7k articles from the LitCovid database is assigned up to seven topics.**

Methods: We propose two different approaches for the multi-label assignment task. The first approach uses the untagged training and validation datasets, while the second approach uses the tagged training and validation datasets. In both cases, the final model is obtained by fine-tuning BioBERT separately on the abstract part of the text and the 'remaining' part of the text consisting of title and metadata. Specifically, for each of the approaches, the training resulted in two fine-tuned BioBERT models, namely, the model trained on the abstracts part and the model trained on the 'remaining' part. For each approach, the final prediction is obtained by performing an ensemble of the prediction outcomes of these two models. We refer to the final models obtained as part of our first approach as Model 1, and the model obtained as part of the second approach as Model 2.

Results: On the test data, Model 1 achieved instance-based f1 score, label-based macro f1 score and label-based micro f1 scores of 0.8845, 0.8495 and 0.7896 respectively. Model 2 achieved scores of 0.8267, 0.8157 and 0.7181 respectively. The baseline model (MLNet) provided by the challenge organizers achieved scores of 0.8678, 0.7655 and 0.8437 respectively.

Conclusions: Our Model 1 showed better performance than Model 2 on both label and instance based F1 scores. Model 1 showed better label-based macro and instance-based F1 scores than the challenge baseline model (MLNet). Further, when benchmarked against the 80 valid submissions for this challenge, Model 1 label-based macro F1-score was close to the median F1 score and instance-based F1-score was close to mean score.

Keywords* — *BioCreative VII, BioBERT, PRIORI-T

I. Introduction

The exponential growth of biomedical literature has significantly increased effort in searching, curating and mining information. As mentioned in the BC-7 Challenge, this was more pronounced with a deluge of pre-prints, journal articles and general COVID-related literature since the beginning of the COVID-19 pandemic. One major resource of COVID literature is the LitCovid curated literature hub [1,2] which contains up-to-date scientific information about the SARS-CoV-2 virus and the pandemic. At the time of writing this article, LitCovid had around 1,84,330 articles and this is rapidly growing. Categorization of these articles based on research topics and geographies are performed for better retrieval.

In order to improve the efficiency of article classification and reduce manual effort, there is a need for developing automated topic annotation methods of the articles in this corpus. Track 5 of BioCreative 7 challenge [3] calls for a community effort to automate topic annotation of these articles from the LitCovid dataset. Specifically, the challenge is a multi-label classification task that assigns one or more labels to an article, where the labels are *Treatment, Mechanism, Prevention, Diagnosis, Transmission, Epidemic Forecasting* and *Case Report*.

II. METHODS

In this section, we describe the two approaches that we developed for this challenge.

a) **LitCovid Preprocessor** : The LitCovid preprocessor reads each article from the LitCovid dataset, splits it into two sections namely Abstract section and Remaining section. The abstract section contains only text from the abstract field while the remaining section comprises of the article title, keywords and journal type metadata fields. The text from the abstract and title fields are normalized to lower case, tokenized using BERT Tokenizer and encoded along with special tokens. Binary encoding was performed for each of the seven labels in the training and validation datasets using a multi-label binarizer.

b) **PRIORI-T Annotator** : The Annotator performs Named Entity Recognition (NER) on the abstract and title texts. Specifically, we repurposed our text-mining framework PRIORI-T [4] to perform NER covering 27 different entity types, namely, human genes, SARS genes, MERS genes, SARS-CoV-2 genes, HPO phenotypes, drugs, chemicals, diseases, disease symptoms, GO process, GO function, GO localization, cell types, tissues, anatomy and non-biomedical entities such as country, non-pharma interventions etc. Conflicts across entity types were resolved by the conflict resolver module. These annotations are then masked by replacing the tagged entity with its entity type token.

c) **BioBERT** : We used the large BioBERT [5] model pretrained on Multi-Genre Natural Language Inference (MNLI) corpus for this challenge. BioBERT-MNLI model was fine-tuned during training. For both the approaches, Adam, Binary Cross Entropy (BCE) with Logits, validation loss were used as optimizer, loss function and checkpoint monitor respectively. The training resulted in two separate fine-tuned BioBERT models, one for the abstract section and one for the remaining section.

d) **Ensemble strategy** : The Ensemble module computes prediction probabilities for labels in an article using the predictions from both the fine-tuned models described above. We used three different ensemble approaches, namely, simple average, weighted average and maximum.

Model 1 : In the training phase of the first approach, we fed the articles in the LitCovid dataset as-is to the LitCovid Preprocessor without performing tagging and entity masking The values of hyperparameters (initial learning rate, batch size, maximum epochs) were 2e-05, 12 and 40 respectively. With regard to the ensemble strategies for this model, maximum score achieved better instance-based and label-based validation F1-scores . Hence we used maximum score ensemble for this approach.

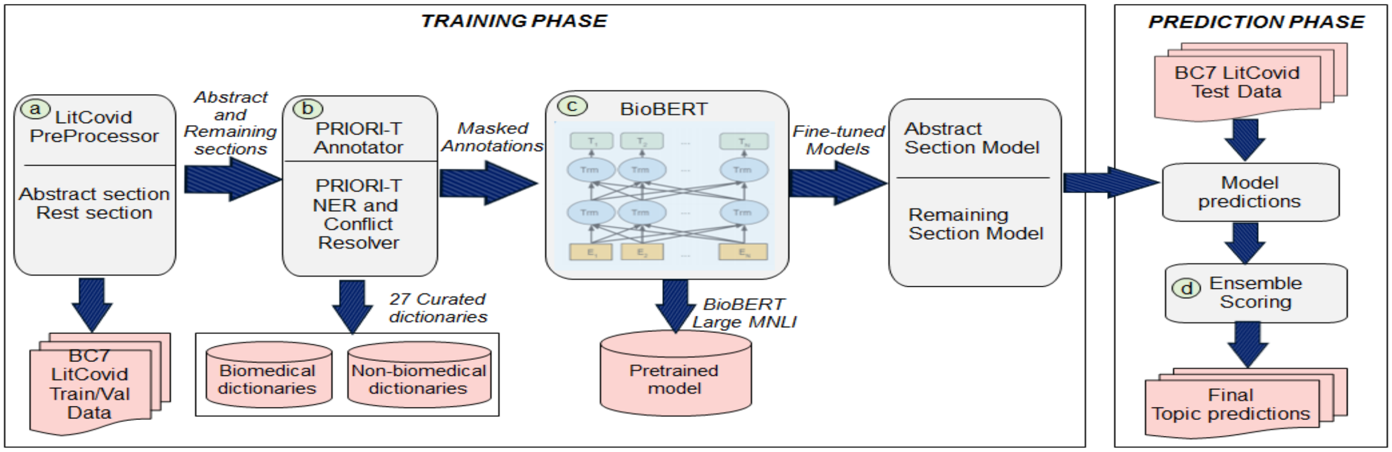

Fig. 1. *Model 2 pipeline*

Model 2 : In the second approach, the output from the LitCovid Preprocessor is input to the PRIORI-T annotator for obtaining masked entity annotations in each of the respective article sections. This was done for train, validation and test data. The hyperparameters and ensemble strategy remained same as that of Model 1. Figure 1 shows the Model 2 pipeline. For Model 1, the only difference is that the PRIORI-T annotator submodule is absent and the output of the preprocessor is directly fed to BioBERT.

III. RESULTS

As shown in Figure 2, Model 1 achieved instance-based F1 score, label-based macro F1 score and label-based micro F1 scores of 0.8845, 0.8495 and 0.7896 respectively. Model 2 achieved scores of 0.8267, 0.8157 and 0.7181 respectively. The baseline model, MLNet[6] provided by the challenge organizers achieved scores of 0.8678, 0.7655 and 0.8437 respectively.

Stats	F1 Scores		
	Label-based macro	***Label-based micro***	***Instance based***
Mean	0.82	0.88	0.89
Std	0.07	0.04	0.05
Q1	0.77	0.85	0.87
Median	0.85	0.89	0.91
Q3	0.87	0.91	0.93
Our Model 1	0.85	0.79	0.89

TABLE I. BC7-TRACK 5 TEAM SUBMISSION STATISTICS

IV. DISCUSSION

Model 1 showed better performance than Model 2 and the challenge baseline model (MLNet). Further, when benchmarked against the 80 valid submissions for this challenge, as shown in Table I, Model 1 label-based macro F1-score was close to the median F1 score and instance-based F1-score was close to mean score. We have also observed class imbalance for topics such as Epidemic forecasting and Transmission which had a low article support in the LitCovid challenge dataset. We tried data augmentation by including additional articles from MEDLINE for these imbalanced classes. However, these augmentations did not improve the performance of both the models.

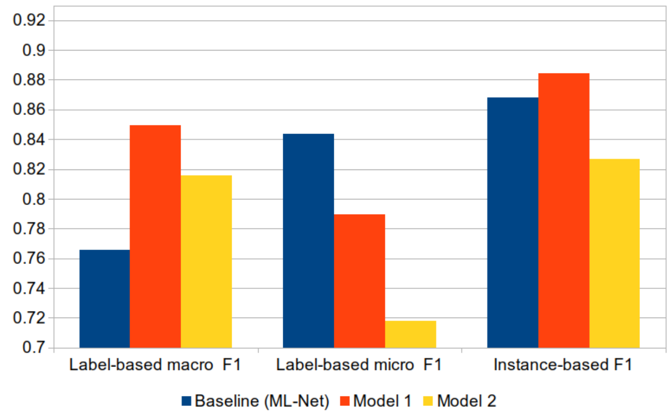

Fig. 2. *Label-based macro, micro and Instance-based F1 scores of Model 1, Model 2 and MLNet on LitCovid Test dataset.*

REFERENCES

1. Chen Q., Allot A., & Lu Z. LitCovid: an open database of COVID-19 literature. Nucleic Acids Research. 2021 Jan 8;49(D1):D1534-40.

2. Chen, Q., Allot, A. and Lu, Z., 2020. Keep up with the latest coronavirus research. Nature, 579(7798), pp.193-194.

3. Qingyu Chen, Alexis Allot, Robert Leaman, Rezarta Islamaj Doğan, and Zhiyong Lu. Overview of the BioCreative VII LitCovid Track: multi-label topic classification for COVID-19 literature annotation. Proceedings of the seventh BioCreative challenge evaluation workshop. 2021.

4. Rao A, Joseph T, Saipradeep VG, Kotte S, Sivadasan N, Srinivasan R., PRIORI-T: A tool for rare disease gene prioritization using MEDLINE. PLoS One. 2020;15(4):e0231728.

5. Lee Ji, Yoon W, Kim S, Kim D, Kim S, So CH and Kang J, BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, Volume 36, Issue 4, 15 February 2020, Pages 1234–1240.

6. Du J, Chen Q, Peng Y, Xiang Y, Tao C and Lu Z. 2019. ML-Net: multi-label classification of biomedical texts with deep neural networks. Journal of the American Medical Informatics Association, 26(11), pp.1279-1285.