Team Bioformer at BioCreative VII LitCovid Track: Multic-label topic classification for COVID-19 literature with a compact BERT model

Li Fang[1], Kai Wang[1, 2*]

[1] Raymond G. Perelman Center for Cellular and Molecular Therapeutics, Children's Hospital of Philadelphia, Philadelphia, PA 19104, USA. [2] Department of Pathology and Laboratory Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA 19104, USA. *Correspondence: wangk@chop.edu

***Abstract*—We describe Bioformer team's participation in the multi-label topic classification task for COVID-19 literature (track 5 of BioCreative VII). Topic classification is performed using different BERT models (BioBERT, PubMedBERT, and Bioformer). We formulate the topic classification task as a sentence pair classification problem where the title is the first sentence, and the abstract is the second sentence. Our results show that Bioformer outperforms BioBERT and PubMedBERT in this task. We also found that further pretraining of Bioformer on COVID-19 articles improves the performance on topic categories of low support. Compared to the baseline results, our best model increased micro, macro and instance-based F1 by 8.8%, 15.5%, 7.4%, respectively.**

Keywords—COVID-19; topic classification; BERT; Bioformer.

I. Introduction

There is an explosion of new scientific literature about the coronavirus disease 2019 (COVID-19). LitCovid is a curated literature resource of COVID-19 studies(1,2). LitCovid is updated daily, and the new articles are curated into eight topic categories. An automated topic classification pipeline can greatly help the curation process. Track 5 of BioCreative VII calls a community effort to develop novel methods for this topic classification problem(3). In this task, each COVID-19-related article can be classified into one or more categories.

Pretrained transformer language models such as BERT (4) and its variants (e.g. RoBERTa(5)) have brought significant performance gains on a variety of language tasks. BERT has been adapted to the biomedical domain (6-8). Recently, we pretrained a compact biomedical BERT model named Bioformer. In this study, we focus on solving the multi-label topic classification problem using Bioformer and other two biomedical BERT models (BioBERT (6) and PubMedBERT (8)). Our results show that Bioformer outperforms BioBERT and PubMedBERT. All the three BERT models provide significant performance increase compared to the baseline methods.

II. Materials and Methods

A. Training, development and test set

The training and development set of the task contain 24960 and 6239 articles, respectively. The test set contains 2500 articles. Each article has the information of journal name, article title, abstract, keywords (optional), publication type, authors, and DOI. Different from the LitCovid website, the task does not include the "General" category and only has seven categories: Mechanism, Transmission, Diagnosis, Treatment, Prevention, Case Report, and Epidemic Forecasting.

B. Models used in this study

We used BioBERT(6), PubMedBERT(8) and Bioformer (https://github.com/WGLab/bioformer/). For BioBERT, we used $\text{BioBERT}_{\text{Base-v1.1}}$, which is the version described in the publication. PubMedBERT has two versions: one version was pre-trained on PubMed abstracts (denoted by $\text{PubMedBERT}_{\text{Ab}}$ in this study), and the other version was pre-trained on PubMed abstracts plus PMC full texts (denoted by $\text{PubMedBERT}_{\text{AbFull}}$). We used $\text{Bioformer}_{\text{8L}}$ which is a compact Biomedical BERT model with 8 hidden layers. $\text{Bioformer}_{\text{8L}}$ was pretrained on PubMed abstracts and one million PMC full-text articles for 2 million steps. We also pretrained $\text{Bioformer}_{\text{8L}}$ on abstracts of COVID-19 articles to see if this leads to a performance gain (described below).

C. Further pretraining Bioformer on COVID-19 articles

We downloaded the abstracts of COVID-19 articles from the LitCovid website. As of Aug 25, 2021, there were 164,179 articles and the total size of the abstracts was 164MB. The pretraining was performed on Google Colab with TPU (v2-8) acceleration. The max input length is fixed to 512 and the batch size was set to 256. The learning rate is set to 2e-5. We pretrained Bioformer on this dataset for 100 epochs with dynamic masking enabled. The number of optimization steps is about 80k. The pretraining was finished in 8 hours. We denote this model as $\text{Bioformer}_{\text{LitCovid}}$.

D. Topic classification

We formulate the topic classification task as a sentence pair classification problem where the title is the first sentence and the

abstract is the second sentence. The input is represented as "[CLS] title [SEP] abstract [SEP]". The representation of the [CLS] token in the last layer was used to classify the relations. We utilized the sentence classifier in transformers python library to fine-tuning the models. We treat each topic independently and fine-tuned seven different models (one per topic). We fine-tuned each BERT model on the training dataset for 3 epochs. The maximum input sequence length was fixed to 512. A batch size of 16 was selected, and a learning rate of 3e-5 was selected.

III. Results

A. Performance on the development set

The performance on the development set is shown in Table I. The performance was evaluated using the script provided by the challenge organizer. As this is a multi-label classification task, four different average F1 scores are presented. Bioformer$_{8L}$ achieves best performance on three metrics: instance-based F1, weighted average F1, and micro F1. Bioformer$_{Litcovid}$ achieves best performance on macro F1. PubMedBERT$_{Ab}$ and PubMedBERT$_{AbFull}$ performers better than BioBERT$_{Base-v1.1}$.

TABLE I. Development Set Performance

Model	Micro F1	Macro F1	Instance-based F1	Weighted average F1
Bioformer$_{8L}$	91.05 (1)	86.60 (3)	91.69 (1)	91.06 (1)
Bioformer$_{Litcovid}$	91.00 (2)	86.64 (1)	91.64 (3)	91.00 (2)
PubMedBERT$_{AbFull}$	90.89 (3)	86.44 (4)	91.68 (2)	90.90 (3)
PubMedBERT$_{Ab}$	90.80 (4)	86.62 (2)	91.49 (4)	90.82 (4)
BioBERT$_{Base-v1.1}$	90.77 (5)	86.14 (5)	91.47 (5)	90.77 (5)

Note: F1 scores are scaled by 100x. The number in the parentheses indicates the ranking of the model.

B. Pretraining of Bioformer on COVID-19 articles improves the performance on topic categories of low support

Macro F1 score is the unweighted mean of each topic category. The above results showed that Bioformer$_{Litcovid}$ has higher macro F1 score than Bioformer$_{8L}$, but micro F1 and weighted F1 score are not better. This indicates that further pretraining of Bioformer$_{8L}$ on COVID-19 abstracts improved the performance on topic categories with fewer support (i.e., number of positive examples). We examined this by comparing the performance on each category (Table II). While the number of articles in the development set is 6239, three topics (transmission, epidemic forecasting, and case report) have less than 500 positive examples, which is a severe imbalance. Bioformer$_{Litcovid}$ has better performance on all the three topic categories.

TABLE II. Performance of Each Category (Development Set)

Topic category	Support	Bioformer$_{8L}$	Bioformer$_{Litcovid}$
Treatment	2207	91.71	91.50 (-0.21)
Diagnosis	1546	89.12	88.97 (-0.15)
Prevention	2750	94.85	95.12 (+0.27)
Mechanism	1073	89.63	88.92 (-0.71)
Transmission	256	72.09	72.12 (+0.03)
Epidemic Forecasting	192	77.52	78.46 (+0.94)
Case Report	482	91.30	91.38 (+0.08)

Note: F1 scores are scaled by 100x. The number in the parentheses indicates the performance improvement compared with Bioformer$_{8L}$.

C. Performance on the test set

We submitted the prediction results of five fine-tuned models (described in Table III). These include three models (Bioformer$_{8L}$, PubMedBERT$_{Ab}$, and BioBERT$_{Base-v1.1}$) fine-tuned on the training set and one model (Bioformer$_{8L}$) fine-tuned the combination of training and development set. The LitCovid website provides more than 164k articles with labeled topics. To test if we can get a performance gain from this information, we fine-tuned Bioformer$_{8L}$ on all labeled articles from the LitCovid website (denoted as Bioformer$_{8L}$-web).

TABLE III. Description of Submitted Models

Fine-tuned model name	Pretrained Model	Fine-tuning data
Bioformer$_{8L}$-train	Bioformer$_{8L}$	training set
PubMedBERT$_{Ab}$-train	PubMedBERT$_{Ab}$	training set
BioBERT$_{Base-v1.1}$-train	BioBERT$_{Base-v1.1}$	training set
Bioformer$_{8L}$-train-dev	Bioformer$_{8L}$	training + dev set
Bioformer$_{8L}$-web	Bioformer$_{8L}$	LitCovid website

The test set results returned by the challenge organizer are shown in Table IV. We also showed the baseline performance(9) and team statistics. We first compare the three models that were fined-tuned on the same dataset (the official training set). Similar to the development set results, Bioformer$_{8L}$ outperforms PubMedBERT$_{Ab}$ and BioBERT$_{Base-v1.1}$ in terms of micro F1 and instance-based F1. PubMedBERT$_{Ab}$ achieved better macro F1 than the other two models. In the development set results, Bioformer$_{Litcovid}$ has a slightly higher macro F1 score than PubMedBERT$_{Ab}$ but we didn't submit the predictions of Bioformer$_{Litcovid}$ due to the limited number of submissions. Fine-tuning on the combination of training and development set improved the micro F1 score, which is often the preferred metric for multi-class classification when there is class imbalance. Fine-tunning on labeled articles from the LitCovid website (Bioformer$_{8L}$-web) failed to improve the performance. After the challenge, we learned that not all articles in the LitCovid website are manually curated. It includes a substantial portion of articles that are classified by text-mining tools. All our submissions provide significant performance gain compared with the baseline method. Our best model (Bioformer$_{8L}$-train-dev) increased micro, macro and instance-based F1 by 8.8%, 15.5%, 7.4%, respectively.

TABLE IV. Test Set Performance

Model	Micro F1	Macro F1	Instance-based F1
Bioformer$_{8L}$-train-dev	91.81 (1)	88.39 (4)	93.24 (2)
Bioformer$_{8L}$-train	91.79 (2)	88.70 (2)	93.34 (1)
BioBERT$_{Base-v1.1}$-train	91.70 (3)	88.63 (3)	93.14 (3)
PubMedBERT$_{Ab}$-train	91.66 (4)	88.75 (1)	93.11 (4)
Bioformer$_{8L}$-web	90.35 (5)	87.43 (5)	91.69 (5)
Baseline (ML-Net)(9)	84.37	76.55	86.78
Mean of all teams	87.78	81.91	89.31
Q1 of all teams	85.41	76.51	86.68
Median of all teams	89.25	85.27	91.32
Q3 of all teams	90.83	86.70	92.54

Note: F1 scores are scaled by 100x. The number in the parentheses indicates the ranking in our five submissions (not the ranking among all teams).

IV. Discussion

In this paper, we present Bioformer team's approaches for the LitCovid Multi-label Topic Classification Track. Our results show that Bioformer outperforms two other BERT models in this task. Our best model provides significant performance gain compared with the baseline method. Predictions for topic categories with low support are more challenging due to lack of training examples. We showed that further pretraining of Bioformer on COVID-19 articles could improve the performance on these categories. We expect that the best results can be achieved by using a combination of Bioformer_{8L} and $\text{Bioformer}_{\text{Litcovid}}$. We hope our study facilitate the topic classification of COVID-19 articles.

References

1. Chen, Q., Allot, A., and Lu, Z. (2021). LitCovid: an open database of COVID-19 literature. Nucleic Acids Res *49*, D1534-D1540. 10.1093/nar/gkaa952.
2. Chen, Q., Allot, A., and Lu, Z. (2020). Keep up with the latest coronavirus research. Nature *579*, 193. 10.1038/d41586-020-00694-1.
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. Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. held in Minneapolis, Minnesota, jun. (Association for Computational Linguistics), pp. 4171-4186.
5. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692.
6. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., and Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics *36*, 1234-1240. 10.1093/bioinformatics/btz682.
7. Peng, Y., Yan, S., and Lu, Z. (2019). Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. held in Florence, Italy, aug. (Association for Computational Linguistics), pp. 58-65.
8. Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., and Poon, H. (2020). Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. arXiv:2007.15779.
9. 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. J Am Med Inform Assoc *26*, 1279-1285. 10.1093/jamia/ocz085.