# Team DonutNLP at BioCreativeVII LitCovid Track: Multi-label Topic Classification for COVID-19 Literature Annotation using the BERT-based Ensemble Learning Approach

Sheng-Jie Lin[1], Yu-Wen Chiu[1], Wen-Chao Yeh[2], Yung-Chun Chang[1,*]
[1]Graduate Institute of Data Science, Taipei Medical University, TMU, Taipei, Taiwan
[2]Institute of Information Systems and Applications, National Tsing Hua University, NTHU, Hsinchu, Taiwan

***Abstract*—Scientific articles on COVID-19 and SARS-CoV-2 are published rapidly since the outbreak of the current pandemic in 2020, with about 10,000 new articles each month. Researchers, healthcare professionals, and even the public are finding it increasingly difficult to stay on top of the latest research. To solve this serious problem, BioCreative holds a LitCovid track that calls for a community effort to tackle the task of automated topic annotation for COVID-19 literature. To this end, we propose a BERT-based ensemble learning system for identifying topics in COVID-19 literatures. In the official results, our method achieves remarkable performances with precision, recall, and F1-score of 94.4%, 92.54%, and 93.46%, respectively.**

***Keywords—BERT; ensemble learning; multi-label classification***

## I. INTRODUCTION

The COVID-19 pandemic has greatly affected our society in a number of aspects, including increased mortality and morbidity, disruption of daily life, and general uncertainty (1). Several epidemiological, clinical, and laboratory studies have unfolded in response to the Coronavirus pandemic, providing policy makers with insights into how to manage the current and future medical and public health issues. In addition, more than 10,000 new articles on the new pandemic are published monthly, which indicates that there has been a rapid increase in the amount of publications on SARS-CoV-2 and COVID-19. They include research exploring the disease, its causes, treatments, etc., with over 187,206 articles in PubMed (2). Consequently, many researchers are suffering from information overload, which makes it difficult to keep up with the latest progress. Therefore, Natural Language Processing (NLP) has been adopted by many for reducing the manual effort on managing biomedical literature in recent years (3).

Chen et al. (2) constructed an open database of COVID-19 literature in 2021, called LitCovid. It has accumulated more than 100,000 articles, with millions of accesses each month by users worldwide. LitCovid is updated daily, and this rapid growth significantly increases the burden of manual curation. In particular, annotating each article with up to eight possible topics, e.g., Treatment and Diagnosis, has been a bottleneck in the LitCovid curation pipeline.

In light of this, BioCreative VII LitCovid Track devotes to tackle automated topic annotation for COVID-19 literature (4). Topic annotation in this competition is a multi-label classification task that assigns one or more labels to each article. These topics have been demonstrated to be effective for information retrieval and have been used in many downstream applications related to LitCovid. However, annotating these topics has been a primary bottleneck for manual curation. Increasing the accuracy of automated topic prediction in COVID-19-related literature would be a timely improvement beneficial to curators and researchers worldwide.

To tackle this task, we employ an ensemble learning method to integrate multiple BERT models (5) for detecting topics in COVID-19 literature. Moreover, we try to fuse linguistic features into BERT in order to boost its prediction performance. The results demonstrate that the proposed system can outperform the benchmark method, median, and Q3 of comparisons, as well as achieve remarkable outcome.

## II. Dataset and Methods

### A. Dataset

This research adopts the articles from LitCovid (2), a literature database of COVID-19-related papers in PubMed that has accumulated more than 100,000 articles. The training and development datasets contain over 30 thousand publicly available COVID-19-related articles and their metadata (e.g., title, abstract, journal). They have been manually reviewed and annotated by in-house models. On the other hand, similar to the training and development datasets, the evaluation set contains 2,500 articles that have been manually reviewed. Each article can contain one or more of eight topics, namely, *Treatment*, *Diagnosis*, *Prevention*, *Mechanism*, *Transmission*, *Epidemic Forecasting*, and *Case Report*.

### B. Methods

To predict the topic(s) of COVID-19 literature, we use a pre-trained BERT model as the basis of our system, and further integrate multiple models through ensemble learning. The

TABLE II. LitCovid track Performance

| Submit ID | **Label-based Micro-avg.** | **Label-based Macro-avg.** | **Instance-based** |
|---|---|---|---|
| | *Precision* / *Recall* / *$F_1$-score* | | |
| Baseline_ ML-Net | 0.8756 / 0.8142 / 0.8437 | 0.8364 / 0.7309 / 0.7655 | 0.8849 / 0.8514 / 0.8678 |
| Mean | 0.8967 / 0.8624 / 0.8778 | 0.8670 / 0.8012 / 0.8191 | 0.8985 / 0.8887 / 0.8931 |
| Q1 | 0.8803 / 0.8452 / 0.8541 | 0.8463 / 0.7545 / 0.7651 | 0.8699 / 0.8619 / 0.8668 |
| Q3 | 0.9251 / 0.8964 / 0.9083 | 0.9079 / 0.8555 / 0.8670 | 0.9353 / 0.9192 / 0.9254 |
| BC7_submission_14 | 0.9343 / **0.9010** / **0.9174** | **0.9214** / 0.8417 / 0.8725 | **0.9440 / 0.9254 / 0.9346** |
| BC7_submission_43 | 0.9350 / 0.8946 / 0.9144 | 0.9152 / **0.8566** / **0.8754** | 0.9459 / 0.9222 / 0.9339 |
| BC7_submission_44 | **0.9395** / 0.8852 / 0.9116 | 0.9047 / 0.8402 / 0.8646 | 0.9462 / 0.9137 / 0.9297 |
| BC7_submission_46 | 0.9311 / 0.8963 / 0.9133 | 0.9084 / 0.8372 / 0.8639 | 0.9426 / 0.9223 / 0.9323 |
| BC7_submission_64 | 0.9342 / 0.8877 / 0.9104 | 0.9157 / 0.8437 / 0.8702 | 0.9457 / 0.9190 / 0.9322 |

different systems we develope for this track are described as follows (Here, we use submission id as the name of a method):

- *BC7_submission_14:* since the BioBERT model is efficient in biomedical NLP tasks (6), we utilize ensemble learning with a majority voting mechanism to integrate multiple BioBERT models, which are selected through k-fold cross validation, for this submission.

- *BC7_submission_43:* in this submission, we investigate the performance improvement by merging different types of BERT models. In addition to BioBERT, the Sultan (7) is the latest BERT model published in 2021 that achieves state-of-the-art performance on several biomedical NLP tasks. The Sultan model is built from the ELECTRA architecture (8) and trained on PubMed abstracts with biomedical domain vocabulary for 434K steps and a batch size of 4096. Therefore, we integrate Sultan with two different versions of BioBERT (v1.1 and v1.2) through a majority voting mechanism.

- *BC7_submission_44*: in this competition, we observe that the dataset is imbalanced, which indicates an unequal distribution of classes within a dataset. To handle the data imbalanced issue, we set the class weights of the final layer of the BioBERT-classifier inversely proportional to their respective frequencies.

- *BC7_submission_46*: we based on *BC7_submission_14* to add extra features from section labels, such as title, abstract, and table, etc. More specifically, we obtain section labels through extracting 'pubtype' of instances in the dataset, and concatenate it with the BERT embeddings.

- *BC7_submission_64:* in this submission, we adopt ensemble learning to merge multiple Sultan models selected by k-fold cross validation.

TABLE I. Process of finding the best Parameter

| **Parameter** | **Test Range** | **Selected** |
|---|---|---|
| Batch size | 8-128 | 32 |
| Learning rate | 1e-5, 2e-5, 3e-5, 4e-5, 5e-5 | 2e-5 |
| Epoch | 4-10 | 6 |
| Weight decay | 1e-3, 2e-3, 3e-3 | 1e-3 |
| Random seed | 5, 42, 9527 | 42 |

## III. RESULT AND DISCUSSION

In this competition, different types of $F_1$-scores are used as evaluation metrics, including: (a) Label-based micro-averaged, a harmonic mean of the per-class $F_1$-score. (b) Label-based macro-averaged, an arithmetic mean of the per-class $F_1$-score. (c) Instance-based, the $F_1$-score calculated by the official script. Our models are developed using PyTorch, a Python deep learning library. The pretrained BERT models are provided by the HuggingFace NLP library. We use 'biobert-base-cased-v1.2' and 'biobertv1.1-pubmed' for different versions of the BioBERT model, and 'BioM-ELECTRA-Large-SQuAD2' for the Sultan model. In addition, we perform hyperparameter tuning as follows. To evaluate a particular set of hyperparameters, the model is trained on the training set using specific hyperparameter values, and its performance is compared with the validation set. Table I provides the test ranges of the hyperparameter tuning process.

The experimental results on the test set are shown in Table II. In the challenge, each team can submit up to five results. The organizers provided team submission-related statistics that contains the mean, standard deviation, Q1, median, and Q3 of the measures for all submissions, and ML-Net (9) serves as the baseline model for comparison. As shown in Table II, all of our models significantly outperform the baseline with an approximately 7% relative increase in instance-based $F_1$-score compared with the best results of the baseline. It is worthy to note that the instanced-based $F_1$-score of our best model outperforms Q3 of the measures for all the submissions by almost 1%. For the results of *BC7_submission_14*, it is shown that utilizing k-fold cross validation as the model selection approach for multiple BioBERT version 1.2 models and further integrating them using a majority voting mechanism is efficient in multi-label topic classification for COVID-19 literature annotation.

## IV. Conclusion

This research presents our methodology and results based on deep learning models for the multi-label topic classification to COVID-19 literature annotation in the BioCreative VII LitCOVID track. In our experiments, using pretrained models are efficient in predicting topics about COVID-19 literature. Moreover, the effect of ensemble learning with majority voting mechanism for the pretrained model is investigated. The outcome suggests that ensemble learning is effective in improving the performance of our models for the LitCOVID track. However, from our error analysis, it is concluded that our models may be further improved by considering additional information such as keywords, which are at a finer scale than pubtypes, and external knowledge bases. These directions will be explored in our future work.

## Acknowledgment

This research was supported by the Ministry of Science and Technology of Taiwan under grant MOST 109-2410-H-038 -012 -MY2.

## References

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

2. Chen, Q., Allot, A., & Lu, Z. (2021). LitCovid: an open database of COVID-19 literature. *Nucleic Acids Res*, 49(D1), D1534-D1540.

3. Chen, Q., Leaman, R., Allot, A., Luo, L., Wei, C. H., Yan, S., & Lu, Z. (2021). Artificial Intelligence in Action: Addressing the COVID-19 Pandemic with Natural Language Processing. Annual *Review of Biomedical Data Science*, 4.

4. 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. In *Proceedings of the seventh BioCreative challenge evaluation workshop*, 2021.

5. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint* arXiv:1810.04805.

6. Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4), 1234-1240.

7. Alrowili, S., & Vijay-Shanker, K. (2021, June). BioM-Transformers: Building Large Biomedical Language Models with BERT, ALBERT and ELECTRA. In *Proceedings of the 20th Workshop on Biomedical Language Processing* (pp. 221-227).

8. Clark, K., Luong, M. T., Le, Q. V., & Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. arXiv preprint arXiv:2003.10555.

9. Du, J., Chen, Q., Peng, Y., Xiang, Y., Tao, C., & Lu, Z. (2019). ML-Net: multi-label classification of biomedical texts with deep neural networks. *J Am Med Inform Assn*, 26(11), 1279-1285.