Team DUT914 at BioCreative VII LitCovid Track：A BioBERT-based feature enhancement approach

Wentai Tang[1], Jian Wang[1§], Hongtong Zhang[1], Xin Wang[1]

[1]College of Computer Science and Technology, Dalian University of Technology, Dalian, 116024, China

***Abstract*—With the rapid expansion of related publications on COVID-19, it becomes more and more important to categorize the available literature automatically in managing massive information. The wide application of pre-trained language model such as BERT and the flexibility of fine-tuning on task-specific data have shown promising results in the field of document classification. For this reason, we have chosen two variants of the transformer model as the basis, namely CovidBERT and BioBERT. Based on the best-performing model BioBERT, we combine the abstract textual representation with a title embedding. Finally, a new information enhancement method was designed to utilize the fusion of label information, which resulted in an instance-based F1 of the model of 93.94% (93.5% precision, 94.38% recall).**

Keywords—Pre-trained language model; COVID-19; Multi-label topic classification; Information enhancement.

I. INTRODUCTION

The rapid growth of biomedical literature poses significant challenges to manual curation and interpretation. This challenge has become more evident during the COVID-19 pandemic(10). Categorizing the available literature plays an important role in managing the flood of information. Unlike the traditional classification tasks, this track focuses on multi-label text classification of documents(4) on COVID-19. Each article in the LitCovid(1) dataset is labeled with one or more up to eight categories, which is a document-level multi-label classification task.

There are several challenges in this document-level multi-label classification task. 1. Since many encoders perform poorly on long text, how to choose an effective encoder for document-level data is worth considering(6); 2. information entropy does not match the text length, and short titles often contain greater information entropy(7); 3. label information is not fully utilized, and label information in the training set is usually ignored(8).

We build a document classification model based on a pre-trained BERT-style model. In addition, to address the mismatch between information entropy and text length in the data, we pre-process the data to balance the information entropy of document titles and abstracts. Finally, we design a new information enhancement method that uses the association between labels. The method incorporates label information from the training set during training.

II. METHODS

We design a BioBERT-based(3) feature enhancement approach which includes three modules as shown in *Fig. 1*. First, text preprocessing. It aims to obtain the input format required by the model. Second, feature processing. We strengthen the weight of short headings in the input module by changing the combination of summary and headings. Thus the entropy of the input information is equalized in each section. Finally, feature enhancement. We use the label distribution of the training set to obtain a label distribution matrix. By introducing this matrix into the model, the label information is incorporated.

A. Text preprocessing

First, we extract the article titles and the abstracts from the dataset. Then the article title and the abstract are concatenated as the first input part. Additional, we only take the article titles as the second input part. Second, we count the distribution of labels in the training set as shown in Table I, and design a tag association matrix based on the distribution of labels.

B. Feature Processing

Feature processing is designed to achieve feature equalization. The first input part which consists of the title and the abstract is tokenized and then encoded by BioBERT. The second input part including only the title is embedded randomly. Then, we concatenate the processed features and the tokenization of the title to obtain the equalized features.

C. Feature Enhancement

We design the feature enhancement module to integrate label features into the model. Previously, we have obtained the label matrix by text preprocessing. Then, we multiply the equalized features by the label matrix to obtain the final output vector used for classification.

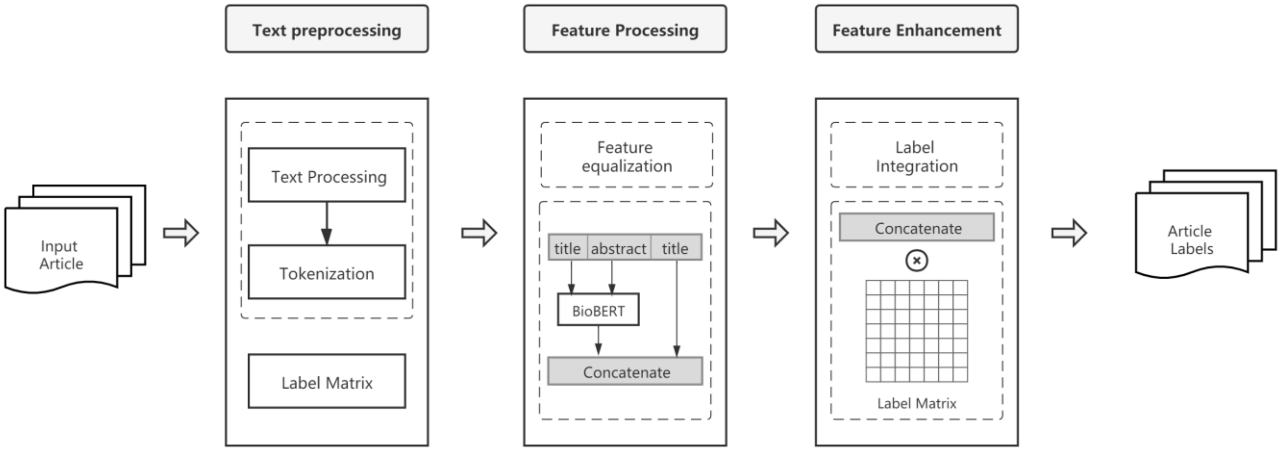

Fig 1.The processing flow of our method

III. EXPERIMENT AND RESULTS

A. *Distribution of text length*

To analyze the data, we count the data distribution of title and abstract lengths, as shown in Fig 2 and 3.

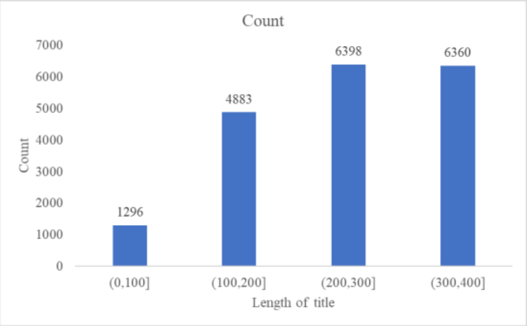

Fig 2. Length distribution of titles

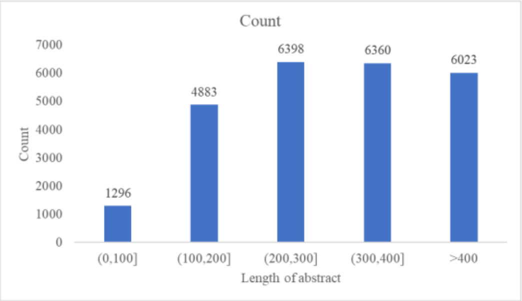

Fig 3. Length distribution of abstracts

As we can see, the length of the title is much less than that of the abstract. However, the amount of information possessed by a title is not proportional to its length. Therefore, we make a treatment for the information imbalance between the title and the abstract.

B. *Label distribution*

In addition, we count the distribution of labels in the training set as shown in Table I.

TABLE I. LABEL RELEVANCE DISTRIBUTION

class	Tre	Dia	Pre	Mec	Tra	Ep-Fore	Case-Re
Tre	1	0.34	0.07	0.39	0.01	0.00	0
Dia	0.48	1	0.11	0.12	0.04	0.00	0
Pre	0.06	0.06	1	0.01	0.05	0.04	0
Mec	0.77	0.16	0.04	1	0.05	0.00	0
Tra	0.12	0.26	0.56	0.23	1	0.06	0
Ep-Fore	0.01	0.01	0.65	0.01	0.09	1	0
Case-Re	0	0	0	0	0	0	1

We obtain the corresponding distribution matrix by counting the distribution of the labels in the training set. Specifically, count the total number of occurrences of the first label, M. Then count the total number of occurrences of both the first and the other label as N. N/M is the correlation value. Due to space limitations, we have abbreviated the names of each category.

C. *Task Results*

Table II shows the results from the task: LitCovid track Multi-label topic classification for COVID-19 literature annotation.

TABLE II. RESULTS FROM THE TASK

model	**precision**	**recall**	**F1**
Baseline (ML-Net)	88.49	85.14	86.78
CovidBERT	90.91	91.65	91.28
BioBERT	91.78	91.73	91.75
BioBERT (Title + Abstract)	92.02	93.65	92.83
Ours	**93.50**	**94.38**	**93.94**

Our baseline model is ML-Net(5) provided by the organizer. To effectively handle the document-level data about Covid-19, we use the pre-trained model CovidBERT(2) with 4.5 points above the baseline model. After comparison, BioBERT has better results on this dataset. Therefore, we chose to improve the method based on BioBERT. Compared to BioBERT only using the abstract, our model was 2.19 points higher. Finally, the F1 of our model is 7.16 points higher compared to the baseline model. Specifically, the results for each class of our method are shown in Table III.

TABLE III. RESULTS FOR EACH TYPE OF MODEL

class	precision	recall	F1
Tre	90.3	92.84	91.55
Dia	86.8	92.3	89.47
Pre	93.76	96.73	95.22
Mec	90.13	88.54	89.33
Tra	67.84	75	71.24
Ep-Fore	75.85	81.77	78.7
Case-Re	88.26	93.57	90.84

D. *Ablation experiments*

To analyze the effectiveness of our method, we conduct additional experiments and the results are shown in Table IV.

TABLE IV. ABLATION EXPERIMENTS

model	precision	recall	F1
BioBERT	91.78	91.73	91.75
BioBERT +Feature Processing	91.93	93.69	92.8
BioBERT +Feature Enhancement	91.44	94.5	92.94
Ours	**93.50**	**94.38**	**93.94**

The above experiments showed that the F1 of the model improved by 1.05 points after adding our feature processing method to the original BioBERT. In addition, the F1 of the model improved by 1.19 points after adding our feature enhancement method. Finally, the F1 of our complete model is 2.19 points higher than the F1 of BioBERT. It demonstrates the effectiveness of the improved approach we have devised for multi-label topic classification of COVID-19.

IV. CONCLUSION

In a multi-label topic classification task for COVID-19 literature annotations on the LitCovid track, we design a feature enhancement approach to address the problem of insufficient features in medical datasets. First, we choose a pre-trained model to process the document-level data. Secondly, the input text is feature processed to balance the information entropy. Finally, we incorporate the label distribution information into the model for feature augmentation. A series of experiments have shown that our approach is indeed effective. The F1 value for the test set returned by the organizer was 93.94%.

REFERENCES

1. Chen, Qingyu, et al. “LitCovid: An Open Database of COVID-19 Literature.” Nucleic Acids Research, vol. 49, 2021.

2. Hebbar, Shashank, and Ying Xie. “CovidBERT-Biomedical Relation Extraction for Covid-19.” The International FLAIRS Conference Proceedings, vol. 34, no. 1, 2021.

3. Lee, Jinhyuk, et al. “BioBERT: A Pre-Trained Biomedical Language Representation Model for Biomedical Text Mining.” Bioinformatics, vol. 36, no. 4, 2019, pp. 1234–1240.

4. Adhikari, Ashutosh, et al. “Rethinking Complex Neural Network Architectures for Document Classification.” Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 4046–4051.

5. 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.

6. Yang, Pengcheng, et al. “SGM: Sequence Generation Model for Multi-Label Classification.” Proceedings of the 27th International Conference on Computational Linguistics, 2018, pp. 3915–3926.

7. Ostendorff, Malte, et al. “Enriching BERT with Knowledge Graph Embeddings for Document Classification.” KONVENS, 2019.

8. Ye, Qinyuan, et al. “Looking Beyond Label Noise: Shifted Label Distribution Matters in Distantly Supervised Relation Extraction.” Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019, pp. 3839–3848.

9. Ostendorff, Malte, et al. “Enriching BERT with Knowledge Graph Embeddings for Document Classification.” KONVENS, 2019.

10. 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.