Team CUNI-NU at BioCreative VII LitCovid Track: Multi-label Topical Classification of Scientific Articles using SPECTER Embeddings with Dual Attention and Label-Wise Attention Network

Aakash Bhatnagar[1], *Nidhir Bhavsar*[1], *Muskaan Singh*[2], *Tirthankar Ghosal*[2]

[1]Navrachana University, Vadodara, India
[2]Institute of Formal and Applied Linguistics,
Faculty of Mathematics and Physics,
Charles University, Czech Republic

(1824526,18103488)@nuv.ac.in, (singh,ghosal)@ufal.mff.cuni.cz

Abstract

Subject-Article classification is an important problem in Scholarly Document Processing to address the huge information overload in the scholarly space. This paper describes the approach of our team CUNI-NU for the Biocreative VII-Track 5 challenge: Litcovid multi-label topic classification for COVID-19 literature [1]. The concerned task aims to automate the manual curation of biomedical articles into seven distinct labels, specifically for the LitCovid data repository. Our best performing model makes use of the SPECTER [2] document embeddings for representing abstract, and titles of scientific articles followed by a Dual-Attention [3] mechanism to perform the multi-label categorization. We achieve significantly better performance than the baseline methods. We make our code available at https://github.com/Nid989/CUNI-NU-Biocreative-Track5

Index Terms: subject-article classification, multi-label topical classification, COVID-19 articles, LitCovid, SPECTER, Dual-Attention, LWAN

1. Introduction

With the COVID-19 pandemic impacting almost all spheres of life, there has been a commendable community-wide effort, especially from the biomedical research community, to tame and tackle the various challenges that the pandemic has posed to humanity. Amongst other research communities, Natural Language Processing (NLP) and Machine Learning (ML) came forward to do their bit. The primary goal of such efforts from NLP/ML were to accelerate knowledge discovery via mechanisms to ease storage and retrieval of COVID-19 articles [4], information extraction from COVID-19 publications [5], finding hidden links between concepts via publication mining [6], etc. One of such efforts is the LitCovid [7][8] repository, a literature database of COVID-19-related papers in PubMed that 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. LitCovid is updated daily, and this rapid growth significantly increases the burden of manual curation. Hence automated mechanisms to categorize COVID-19 articles into a set of pre-defined topics is an important use case for LitCovid. There are several categories of articles in LitCovid according to their scope: *Forecasting, Transmission, Mechanism, Case Report, Diagnosis, Treatment, Prevention*. Each category of articles caters to the different information needs of the user. Again, one scientific article may belong to more than one topic, hence could warrant multiple labels. The organizers of the BioCreative VII LitCovid track[1] drew community attention to this curation problem and introduced a challenge for multi-label topical classification of COVID-19 articles in LitCovid.

This paper documents our participation in this challenge and reports our system description and performance on the task. We experimented with several textual representations and neural architectures. In our attempts, we performed the best when we used SPECTER document embeddings [2] with dual attention on a label-wise attention network [9]. The main advantage of using SPECTER over other pre-trained models for this task is that SPECTER has been trained on scientific articles and probably comprehends scientific discourse better than generic models trained on other texts.

2. Dataset and Data Preprocessing

The Litcovid dataset includes the following fields: *pmid (PubMed Identifier), journal, title, abstract, keywords, pub_type, authors, DOI (Digital Object Identifier) and the labels* of publicly available articles. The training and the development data contain 24,960 and 6,239 instances, respectively. As we mentioned earlier, the dataset consists of articles from seven categories: Case-Report, Prevention, Treatment, Transmission, Forecasting, Mechanism, and Diagnosis. The distribution of articles for these labels is in Table 1. The dataset contains far more articles with Prevention and Treatment labels, thus making it a class imbalanced one. We additionally scraped approximately

[1]https://biocreative.bioinformatics.udel.edu/events/biocreative-vii/biocreative-vii/

4,000 unique articles for the Transmission and Forecasting categories from the Litcovid repository to address the data imbalance problem.,

Label	No. of articles
Forecasting	461
Transmission	1,065
Mechanism	3,549
Case Report	1,914
Diagnosis	4,754
Treatment	6,897
Prevention	11,042

Table 1: *Label wise distribution of articles on Original Litcovid dataset*

3. Methodology

As mentioned earlier, we use the SPECTER representations for this task, which produces the document-level embedding using citation-based transformers. Furthermore, the SPECTER model incorporates SciBERT, making it the ideal model for representing the dense bio-medical vocabulary present in the COVID-19 literature. We format the input sentence as below to generate suitable sentence embeddings.

$$[CLS]\ \ title\ \ [SEP]\ \ abstract\ \ [CLS]$$

Next, we use a Dual-attention module, which consists of two self-attention[10] processes that are applied to the embeddings in sequential order. These self-attention[10] layers allow each input to establish relationships with other instances. To obtain unique vectors, i.e., query (Q), key (K), and value (V), three individually learned matrices are multiplied with the input vector. Each of these vectors is associated with R^d and is used to compare a word w_i to every other word in the sentence. This is accomplished using the approach outlined below, in which we apply the dot product to the query and key vector and normalize them using the square root of the key vector's dimension.

$$f\left(Q, K_i\right) = \begin{cases} Q^r K_i & \textit{dot} \\ Q^T W_a K_i & \textit{general} \\ W_o' \left[Q : K_i\right] & \textit{concat} \\ v_s^T \tanh\left(W_a Q + U_a K_i\right) & \textit{perceptron} \end{cases}$$

The value vector is multiplied with the output generated by applying the softmax function to the derived scores. Furthermore, a given word, the sum of these weighted value vectors produces a self-attention[10] output.

$$\text{Attention}(Q, K, V) = \sum_i a_i V_i \tag{1}$$

However, one limitation of using double self-attention is that it can only generate relationships amongst the input instances while completely discarding the output. The double self-attention mechanism helps retain more information contained in the sentence and thus generate a more representative feature vector for the sentence. To improve the results further and overcome the limitation of dual-attention, we use Label-Wise-Attention-Network (LWAN), which provides attention for each label in the dataset.

LWAN architecture is responsible for improving individual word predictability by paying particular attention to the output labels. It uses an attention-mechanism-like [11] strategy to allow the model to focus on specific words in the input rather than memorizing all of the essential features in a fixed-length vector. Attention for LWAN is calculated as follows:

$$z_{i,l} = w_{a,l} h_i + b_{a,l} \tag{2}$$

$$\alpha_{i,l} = \frac{e^{z_{i,l}}}{\sum_{j=1}^{N} e^{z_{j,l}}} \tag{3}$$

$$s_l = \sum_{j=1}^{N} \alpha_{j,l} h_j \tag{4}$$

$$\beta_l = w_{f,l} s_l + b_{f,l} \tag{5}$$

$$p_l = \frac{e^{\beta_l}}{\sum_{r=1}^{L} e^{\beta_r}} \tag{6}$$

The label-wise attention mechanism generally applies the same attention procedure but repeats it L (number of labels) times, where each attention module is reserved for a specific label l.

We used a weighted binary cross-entropy loss to give equal importance to the different classes during the training period, which was necessary due to the large data imbalance. Figure1 explains our overall architecture.

3.1. Other Attempted Methods

We tried several approaches, but the model with SPECTER representation followed by dual-attention and a label-wise attention network performed the best. We experiment with DUAL BERT architecture, which entails fine-tuning two separate pre-trained models to generate output embeddings. These two models have similar architectures, but they were trained on different input sentences. These inputs are then encoded in parallel to produce R^d-shaped sentence embeddings. The final output vector is a weighted average of the sentence embeddings produced by the two employed models.

We use the following pre-trained models during experimentation:

1. SPECTER, a language model to generate document-level embedding of documents.[2]
2. PubMedBERT[12], a biomedical domain-specific BERT model, trained from scratch on PubMed articles[3].
3. COVID-SciBERT[13], a small language modeling expansion of SciBERT, a BERT model trained on scientific text[4].

[2] allenai/specter

[3] microsoft/BiomedNLP-PubMedBERT-abstract

[4] lordtt13/COVID-SciBERT

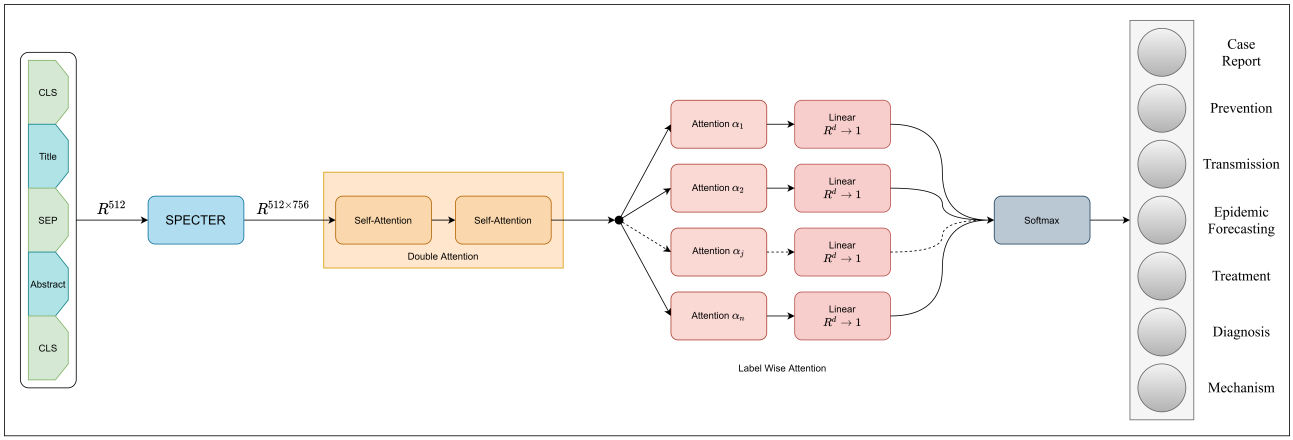

Figure 1: *The architecture of our proposed system for Biocreative VIII Track V*

We prioritize these language models since they are trained for biomedical tasks and can represent these texts better than the other BERT models and their derivatives. Each of these specified models accepts input sequence in the form of input_ids and attention_mask to generate appropriate outputs. Usually, input_ids are the only required parameters to be passed as input. These input_ids are the token indices, which are numerical representations of tokens building the sequences that will be the input for the model. While attention_mask is an optional argument, it is used for batching sequences and indicates which tokens the model should attend. Each input sequence should be padded or truncated to $N = 512$, the maximum value these models can accept. A transformer model constitutes several similar layers stacked on top of each other. Each of these layers accepts input and produces appropriate output. Generally, the output of a BERT model is determined by its last-hidden state, which usually computes a vector ($h_i \in R^{768}$ specific to i^{th} word token. Moreover, since these three models also obey the same architecture, thus producing an output of $R^{512\times 768}$. Table 2 reports our performance with the different models we tried.

Models	Micro-F1	Macro-F1	Instance-F1
PubMedBERT-LWAN	0.8652	0.8047	0.8988
DualBERT-Ensemble	0.8889	0.8570	0.9124
SPECTER-LWAN	0.8871	0.8478	0.8956
CovidSciBERT-LWAN	0.8616	0.8409	0.8576
ML-NET (baseline)	0.7655	0.8437	0.8678
Proposed Method	**0.8959**	**0.8673**	**0.9153**

Table 2: *Label based micro f1, Label based macro f1, Instance based f1 scores for submitted model @ Biocreative VII Track V*

4. Experimental Setup

This section details our experimental setup. We conduct our experiments with Python 3.8 on a GPU Tesla P-100 with RAM 16.28GB on Google Colab. We enlist the hyperparameter setup of our different models in Table 5.

In the Litcovid dataset, there is a significant imbalance of article's labels, particularly for Transmission and Forecasting. We adopted the weighted binary cross-entropy (BCELoss) loss function, which helped solve this problem to some extent. This loss function is used in each of our proposed models, allowing the model to focus on minority classes and improve performance. Here we assign a weight for each class based on the number of instances. Equation 7 formulates the assignment of weights by giving higher values to minority classes than the most dominant class. Here c_i represents the instance count of i^{th} label, c_m is the count of the dominant class and w_i is the associated weight. As shown in Table 3, the weights for Transmission and Forecasting labels are higher as compared to other classes.

$$w_i = \frac{c_m}{c_i} \tag{7}$$

category	value
Case Report	6.1099
Diagnosis	1.8774
Forecasting	9.1609
Mechanism	2.5855
Prevention	1.0
Transmission	4.4227
Treatment	1.3871

Table 3: *Assigned weights to each class*

However, applying weighted BCELoss improved our results marginally from the baseline. As we have seen in Table 2 all of our attempts at modeling a distribution were better than the baseline. The most significant improvement occurred due to the employment of the SPECTER model to generate document embeddings. By embracing the use of citation graphs, SPECTER has been efficiently pre-trained to generate optimal

document embeddings. As a result, SPECTER is a good option for multi-label classifiers since it can create inter-document associations. This explains why other models, even ones pretrained on biomedical domain tasks, could not perform well.

We used a learning rate scheduler which improved our model's performance. For this, we employed a get_linear_schedule_with_warmup scheduler[5]. Thus while training, the learning rate grows to a fixed value of 2×10^{-5} and then goes down linearly to 0. Furthermore, we experimented with multiple values of batch size, ranging from 4-16.

5. Results and Analysis

Our model outperformed the baseline model by a significant margin, particularly for the labels: Case Report, Forecasting, Transmission, and Diagnosis, where the F1 score increased by +8–10 points. Figure 3 shows the F1-score for each label based on the number of annotations to the articles. In the dual label scenario, it is clear that the model is performing better. The results were evaluated over two principal matrices: Macro and Micro averages. Macro-average calculates the metric independently for each class and then averages the results, treating all classes equally. In contrast, a micro-average calculates the average metric by aggregating each class' contribution. Furthermore, because macro-average treats each class equally, it places a greater emphasis on rare classes. As shown in Figure 2, our model performs significantly better for macro-average F1 than other submissions.

As illustrated in Figure 2, our model performed well for Label-based macro F1 and ranked in the 3^{rd} quartile for Label-based micro F1 and Instance-based F1. Figure 2 also shows the mean of all participants for each metric, from where we can infer that our model performed better than most submissions.

Overall, F1 scores are vastly different for uni-label and multi-label classes, especially in Transmission and Mechanism, where the F1-score fluctuates substantially. One possible explanation for this fluctuation is that the dataset contains many articles with *mechanisms* and *transmission* classified together or with *prevention*. Another reason might be a considerable correlation between the few categories of the dataset, especially Forecasting, Prevention, and Transmission. Even though these categories are semantically related and some overlap exists, the Transmission and Forecasting tags are predicted in conjunction with the Prevention tag much more frequently than observed in the labels.

Due to the abundance of introductory lines relating to COVID-19 in these articles, we found that models have trouble detecting discriminative regions of the content. As a result, the model cannot distinguish between the relevance of introductory parts and key sentences such as thesis statements and article titles. Furthermore, as previously stated, due to semantic overlap in categories such as Epidemic Forecasting, Prevention, and Transmission, articles with Transmission and Forecasting tags appear to be in conjunction with the Prevention tag during prediction (See Table 4).

[5] transformers/get_linear_schedule_with_warmup

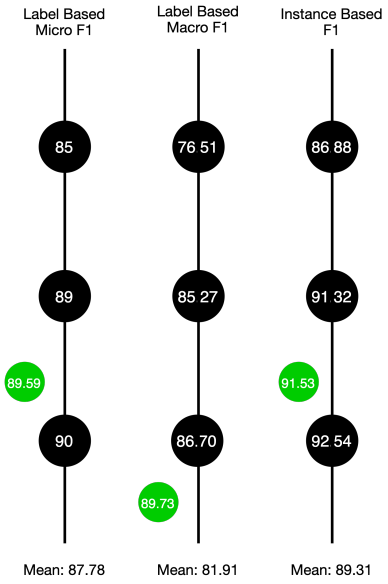

Figure 2: *Position of Team CUNI-NU result as compared to the other participants. The black circles represent the values of Quartile 1, Median, and Quartile 3 respectively. The green circle represents the standing of our model. The range for each line is 0-100.*

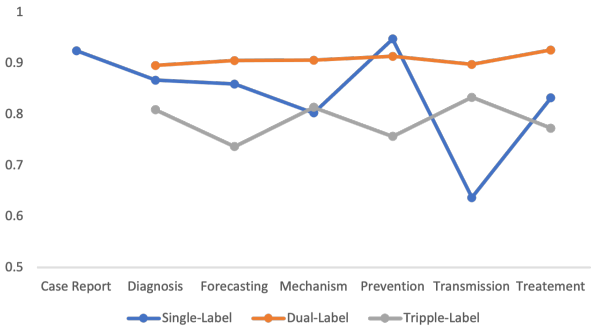

Figure 3: *micro F1 scores for each label in the LitCovid dataset, based on the number of labels assigned to articles instances i.e. single, dual and triple label annotations*

6. Conclusion

We provide an initial method for the Multi-Label classification of COVID-19 literature. Our approach will help reduce the time consumed during manual curation of articles in the Litcovid data repository. We create a pipeline that uses SPECTER embeddings, dual-attention mechanism, and Label-Wise-Attention method for the purpose. Through this approach, we were able to achieve results significantly better than the baselines. We noticed that many abstracts had introductory sentences which were not unique to their label (See Table 4). A method for removing these sentences can enhance the model's performance and

can be explored further. Furthermore, the imbalance in classes should be mitigated to achieve better results.

7. References

[1] S. Tian and J. Zhang, "Multi-label topic classification for covid-19 literature annotation using an ensemble model based on pubmedbert," 2021.

[2] A. Cohan, S. Feldman, I. Beltagy, D. Downey, and D. S. Weld, "Specter: Document-level representation learning using citation-informed transformers," 2020.

[3] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," 2019.

[4] L. L. Wang, K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Burdick, D. Eide, K. Funk, Y. Katsis, R. M. Kinney, Y. Li, Z. Liu, W. Merrill, P. Mooney, D. A. Murdick, D. Rishi, J. Sheehan, Z. Shen, B. Stilson, A. D. Wade, K. Wang, N. X. R. Wang, C. Wilhelm, B. Xie, D. M. Raymond, D. S. Weld, O. Etzioni, and S. Kohlmeier, "CORD-19: The COVID-19 open research dataset," in *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*. Online: Association for Computational Linguistics, Jul. 2020. [Online]. Available: https://aclanthology.org/2020.nlpcovid19-acl.1

[5] D. Das, Y. Katyal, J. Verma, S. Dubey, A. Singh, K. Agarwal, S. Bhaduri, and R. Ranjan, "Information retrieval and extraction on COVID-19 clinical articles using graph community detection and Bio-BERT embeddings," in *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*. Online: Association for Computational Linguistics, Jul. 2020. [Online]. Available: https://aclanthology.org/2020.nlpcovid19-acl.7

[6] D. Herrmannova, R. Kannan, S.-H. Lim, and T. E. Potok, "Covid-19 knowledge graph – dataset for smcdc 2021 challenge 2."

[7] Q. Chen, A. Allot, and Z. Lu, "Keep up with the latest coronavirus research," *Nature*, vol. 579, no. 7798, p. 193, 2020. [Online]. Available: https://www.ncbi.nlm.nih.gov/pubmed/32157233

[8] Q. yu Chen, A. Allot, and Z. Lu, "Litcovid: an open database of covid-19 literature," *Nucleic Acids Research*, vol. 49, pp. D1534 – D1540, 2021.

[9] F. Barbieri, L. Espinosa-Anke, J. Camacho-Collados, S. Schockaert, and H. Saggion, "Interpretable emoji prediction via label-wise attention LSTMs," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018. [Online]. Available: https://aclanthology.org/D18-1508

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017.

[11] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2016.

[12] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-specific language model pretraining for biomedical natural language processing," *CoRR*, vol. abs/2007.15779, 2020. [Online]. Available: https://arxiv.org/abs/2007.15779

[13] I. Beltagy, K. Lo, and A. Cohan, "Scibert: A pretrained language model for scientific text," 2019.

8. Appendix

Article	predicted	true
novel coronavirus disease 2019 COVID-19 caused severe acute respiratory syndrome coronavirus 2 SARS-CoV-2 caused pandemic , threatening global public health current paper , ... multidisciplinary therapeutic approach order achieve favorable clinical outcome , enhancing capability COVID-19 diagnosis use chest imaging modality discussed.	Forecasting, Prevention	Forecasting
Decentralisation decision-making central lower level organisation proposed ... conclusion Decentralisation create condition support innovation improvement locally ... researcher provide actionable knowledge change organisation management could address current challenge healthcare	Forecasting	Forecasting

Table 4: *Example 1 shows that the model predicts prevention along with the forecasting label due to general introductory sentences. These introductory sentences do not provide any helpful information regarding the forecasting label, resulting in incorrect classification. On the other hand, the second example shows that when the abstract does not contain these introductory sentences, the model is predicting the accurate Label*

Model	Hyperparameters	Number of parameters
SPECTER-DualAtt-LWAN	learning rate: 2×10^{-5} max sequence length: 512 batch size: 4 epochs: 10 model: SPECTER warmup proportion: 0.2 dropout probability: 0.1	111M
PubMedBERT-LWAN	learning rate: 2×10^{-5} max sequence length: 512 batch size: 8 epochs: 10 model: PubMedBERT warmup proportion: 0.2 dropout probability: 0.1	$109M$
DualBERT-Ensemble	learning rate: 2×10^{-5} max sequence length: 512 batch size: 8 epochs: 10 model: PubMedBERT warmup proportion: 0.2 dropout probability: 0.1	$220M$
SPECTER-LWAN	learning rate: 2×10^{-5} max sequence length: 512 batch size: 8 epochs: 10 model: SPECTER warmup proportion: 0.2 dropout probability: 0.1	$109M$
CovidSciBERT-LWAN	learning rate: 2×10^{-5} max sequence length: 512 batch size: 8 epochs: 10 model: CovidSciBERT warmup proportion: 0.2 dropout probability: 0.1	$109M$

Table 5: *Hyperparameter Details of our Attempted Systems*