Team PolyU-CBSNLP at BioCreative-VII LitCovid Track: Ensemble Learning for COVID-19 Multilabel Classification

Jinghang Gu[1], Xing Wang[2], Emmanuele Chersoni[1], Chu-Ren Huang[1]

1. Chinese and Bilingual Studies, Hong Kong Polytechnic University, Hong Kong, China;
2. Tencent AI Lab, Shenzhen, China

***Abstract*—This paper briefly describes our works for the LitCovid shared task of BioCreative-VII Track 5. It is an ensemble learning-based system that utilized multiple biomedical pre-trained models. In particular, we leveraged seven advanced models for initialization with homogeneous and heterogenous structures through an ensemble bagging manner. To enhance the representation abilities, we further proposed to employ additional biomedical knowledge to facilitate ensemble learning. The experimental results on the LitCovid datasets show the effectiveness of our proposed approach.**

Keywords—COVID-19; LitCovid; Pre-trained Model; Deep Learning; Multilabel Classification; Ensemble Learning

I. Introduction

Under the scenario of the COVID-19 pandemic sweeping across all over the world, the challenge of the coronavirus has rapidly accelerated the worldwide pace of scientific publications (1,2). Since more than 10,000 articles related to SARS-CoV-2 and COVID-19 would be monthly published (3,4), this rapid growth significantly increases the burden of manual curation. How to precisely curate and interpret this large number of COVID-19 literature has consequently become of great importance for facilitating coronavirus knowledge discovery, clinical prevention, and treatment strategies (5-7).

The identification of semantic topics such as mechanism and treatment from biomedical literature could be helpful for COVID-19 curation. However, manual annotation of such semantic labels from unstructured free text is costly and insufficient to keep up to date. Although some previous attempts (8-11) have been conducted on the free-text datasets, automatic identification of COVID-19 semantic topics still remains challenging. In addition, few identification tools are freely available, and there are limited examples of such tools in real-world applications.

To this end, the BioCreative-VII community proposed a challenging task of LitCovid Track for COVID-19 literature ([12]), which is a standard multi-label classification task that requires each participant to assign one or more semantic labels to each biomedical article. This task was aimed at calling for a community effort to tackle the automated topic annotation for COVID-19 related literature and at providing practical benefits to worldwide biomedical curation.

In this paper, we present the system developed by the PolyU CBS-NLP team for the challenging competition. Our primary goal was to develop a versatile machine learning approach with good robustness and generalizability so as to be easily applied to the COVID-19 domain and scaled up to other biomedical domains. Specifically, we proposed an ensemble learning architecture which leveraged multiple state-of-the-art pre-trained models to address the challenging multilabel classification problem. We extensively explored seven different pre-trained models with homogeneous and heterogenous structures for ensemble learning, guaranteeing the diversity and robustness of these deep neural networks. Moreover, we also proposed to employ extra biomedical knowledge to enhance the semantic representations for the ensemble learning method. The experimental results on the LitCovid datasets show the effectiveness and success of our proposed approach.

II. Dataset

In this section, we first present a brief introduction about the LitCovid corpus, we then systematically depict the statistics of the corpus.

The corpus used for BioCreative-VII LitCovid Track was originated from the LitCovid database (3,4), whose curated data is publicly available for both research discovery and machine processing. More specifically, the organizers collected around 30,000 COVID-19 related articles from the database, which were further split into three subsets of training, development, and test datasets. Since the LitCovid corpus is targeting the multilabel classification for COVID-19, seven topic labels are annotated in the corpus, i.e., *Treatment*, *Diagnosis*, *Prevention*, *Mechanism*, *Transmission*, *Epidemic Forecasting*, and *Case Report*. In addition, the organizers also provided to each article various kinds of metadata retrieved from PubMed.

During the competition phase, the organizers first released the training dataset as well as the development dataset in CSV format. Later, they released the test dataset following the same data format except for the topic labels which are supposed to be predicted by the participants. Detailed information on the LitCovid corpus is shown in Table I and Table II as follows.

Table I presents the basic statistical information of metadata for the corpus. As shown in the table, there are in total 33,699 COVID-19 related biomedical articles collected in the corpus, with a training set of 24,960 articles, a development set of 6,239 articles, and a test set of 2,500 articles, respectively. Most of the articles are filled with valid contents of titles, abstracts, journal names, PMIDs, author names, DOIs, as well as publication types. However, it is worth noting that despite the organizers trying their best to fill the metadata attributes, around 25% of keywords are still missing due to the incompleteness of the online information.

TABLE I. THE METADATA STATISTICS OF THE LITCOVID CORPUS.

Metadata	*Train*	*Development*	*Test*
Title	24,960	6,239	2,500
Abstract	24,900	6,219	2,485
Journal Name	24,960	6,239	2,500
Keywords	18,968	4,754	2,056
PMID	24,960	6,239	2,500
Authors	24,859	6,212	2,499
DOI	24,406	6,100	2,474
Publication Type	24,960	6,239	2,500

Table II depicts the label distribution of the LitCovid corpus. Since the label annotation of the test dataset by far is still unknown, Table II only focuses on the statistical information of the training and development datasets. As shown in the table, it is observed that the frequency of different labels varies a lot. Actually, among all topic labels, the label of *Prevention* dominates the entire corpus with the highest frequency while the label of *Epidemic Forecasting* occurs rarely, indicating an extremely imbalanced label distribution in the corpus, which makes the challenge of the COVID-19 multilabel classification problem even more difficult.

TABLE II. THE LABEL DISTRIBUTION OF THE LITCOVID CORPUS.

Label	*Train*	*Development*
Treatment	8,718	2,207
Diagnosis	6,193	1,546
Prevention	11,102	2,750
Mechanism	4,439	1,073
Transmission	1,088	256
Epidemic Forecasting	645	192
Case Report	2,063	482

III. METHODS

In this paper, an effective ensemble learning paradigm is proposed for COVID-19 multilabel classification. Fig. 1 illustrates the architecture of the proposed method, which is a universal ensemble learning framework integrating multiple classifiers generated from different pre-trained neural models. As known in ensemble learning theory, every single model would be taken as a weak learner or classifier due to its bias and variance in feature representation (13-15). Therefore, the basic idea of our ensemble learning method is to train multiple weak classifiers through an ensemble manner and to aggregate these weak classifiers to acquire better results. Technically, we take multiple state-of-the-art pre-trained neural models as the initialization of the classifiers for the proposed ensemble learning method. Our main hypothesis is that when weak classifiers are correctly aggregated, the system is able to efficiently reduce the bias and variance of such weak learners to create a stronger learner, finally achieving a more accurate and robust performance.

In Fig. 1, each classifier of the pre-trained neural models is first fine-tuned independently during the training process, then all the outputs of these classifiers are merged through an ensemble bagging strategy to obtain the final prediction for COVID-19 multilabel classification. Moreover, in order to improve the representation diversity and robustness of ensemble learning, the pre-trained models with different architectural implementations are mainly considered. Particularly, seven advanced pre-trained models are employed in our ensemble learning method, i.e., BioBERT-Base (16), BioBERT-Large (16), PubMedBERT (17), CovidBERT (18), BioELECTRA (19), BioM-ELECTRA (20), and BioMed-RoBERTa (21). Among these models, there are four variants of BERT (22), two variants of ELECTRA (23), and one version of RoBERTa (24). We refer to all models with the same underlying architecture as homogeneous models, otherwise, we refer to them as heterogeneous models.

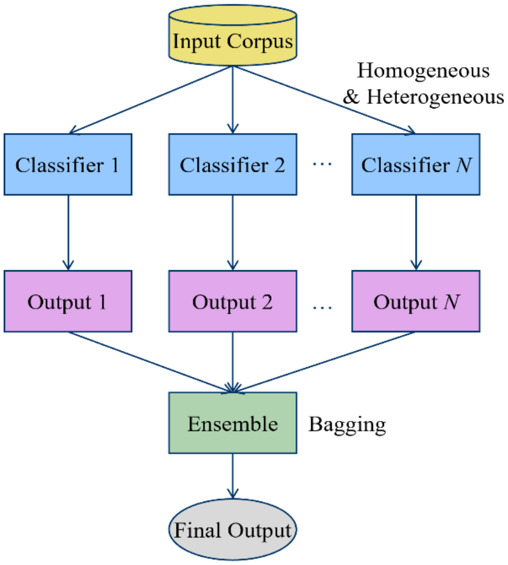

Fig. 1. The ensemble learning framework.

Apart from proposing the ensemble learning method, we also argue that additional biomedical knowledge such as keywords, journals, and MeSH terms, are indispensable for the problem of COVID-19 multilabel classification. The main idea accounting for taking this additional biomedical knowledge is that these kinds of knowledge carry an amount of manually refined semantic information that has been carefully reviewed by authors or curators.

Therefore, before the training process, the input sequence for each article needs to be constructed by concatenating the texts of keywords, journal name, MeSH terms as well as the title and abstract. Note that the MeSH terms are not available in the LitCovid corpus, we thus crawled these textual complements from PubMed, in terms of the corresponding metadata *PMID* of each target article.

IV. EXPERIMENTAL RESULTS

In this section, we first introduce the evaluation methods as well as the experimental settings for the LitCovid track. Then we systematically evaluate the performance of our approach and compare it with the official baselines. The error analysis is finally conducted at the end of this section.

TABLE III. THE PERFORMANCE OF ADDITIONAL BIOMEDICAL KNOWLEDGE ON THE DEVELOPMENT DATASET.

Model	*Knowledge*	*EBP(%)*	*EBR(%)*	*EBF(%)*
BioBERT-Base	title + abstract + keywords + MeSH + journal name	91.02	**90.87**	**90.94**
	title + abstract + keywords + MeSH	91.19	90.65	90.92
	title + abstract + keywords	**91.31**	90.45	90.88
	title + abstract	90.83	88.89	89.85

A. Evaluation Methods and Experimental Settings

Amid the LitCovid evaluation phase, all submissions will be evaluated using both label-based and instance-based metrics that are commonly applied for multi-label classification. Briefly, nine different measurements from three different levels are utilized, i.e., *Macro-Precision (MaP), Macro-Recall (MaR), Macro-F1 (MaF), Micro-Precision (MiP), Micro-Recall (MiR), Micro-F1 (MiF), Example-Based Precision (EBP), Example-Based Recall (EBR),* and *Example-Based F1 (EBF).*

Macro-based measurements are utilized to evaluate the system performance by labels at the macro level. In macro measurements, all the labels are treated equally regardless of their distribution. Correspondingly, micro-based measurements are adopted to evaluate the system performance by labels at the micro level. In micro measurements, the distribution of each label is taken into consideration, and the labels with larger numbers have more impact on the final results during the calculation. In a word, both micro and macro measurements are employed to evaluate the final results reflecting different aspects at the label level. Different from the label-level evaluation, example-based measurements are utilized to evaluate the system performance from another perspective, at the instance level.

In our experiments, all texts of articles and additional biomedical knowledge are converted into lower cases. The default settings of pre-trained models are used for initialization. In the training phase, the binary cross-entropy loss is utilized, and the AdamW optimizer (25) is adopted to minimize the training loss and optimize the model parameters. The learning rates are kept identically for all models with the value of 0.00002 during the training process.

B. System Performance on The Development Dataset

To investigate the importance of the contributions of the proposed biomedical knowledge, we perform the knowledge combination studies on the development dataset trying to reveal the different influences of the knowledge. As seven advanced pre-trained models are utilized for our ensemble learning, the performance with the most naïve model of BioBERT-Base (16) at the instance level is selected to simplify the comparison. Table III exhibits the details of the knowledge combination experiments, in which the best scores are highlighted in boldface. It is worth mentioning that all experiments rely on the basis of contextual information of both titles and abstracts, which are available for all kinds of trials.

In general, it is observed from the table that, merely using the contextual information of titles and abstracts, BioBERT-Base is able to achieve the EBF as high as 89.85%. This suggests that the contexts of biomedical articles contain crucial clues for COVID-19 multilabel classification, and that the pre-trained model can also effectively represent and capture this information. When successively combing the additional biomedical knowledge of keywords, MeSH, and journal name, the performance of the model increases consistently. After the combination of all proposed biomedical knowledge, the performance is further improved up to the EBF score of 90.94%. This indicates that modeling the additional biomedical knowledge is capable of bringing complementary information to the article contexts, which is also helpful for the problem of COVID-19 multilabel classification.

Table IV summarizes the performance of different labels with BioBERT-Base on the development dataset. As shown in the table, the label of *Prevention* achieves the best performance resulting in the highest F1 score of 93.79%. Meanwhile, the labels of *Treatment*, *Diagnosis*, *Mechanism*, and *Case Report* all obtain comparable performances. In contrast, compared with the above labels, labels of *Epidemic Forecasting* and *Transmission* perform the worst. This is probably because of the imbalanced label distribution described in Section *DATASET*, which implies that the fewer class examples, the more difficulties the model will encounter during prediction.

Likewise, Table V demonstrates the comparison at the instance level on the development dataset with all pre-trained models utilized for ensemble learning. It can be observed that all models acquire competitive performance within slight differences due to their state-of-the-art feature representation capabilities. It also indicates that all pre-trained models are able to provide robust COVID-19 specific feature representations which can benefit the ultimate multilabel classification performance. In particular, BioM-ELECTRA (20) has the highest example-based precision while PubMedBERT (17) shows the highest example-based recall. Among all models, BioMed_RoBERTa (21) reports the most advanced performance in example-based F1, achieving the highest score of 92.65%.

Moreover, it is also noticeable that the homogeneous models sometimes rival each other, while heterogenous models always would be diverse. Specifically, the homogeneous variants of BioBERT (16) show similar achievements, similarly to the homogeneous variants of ELECTRA (23). However, when we compare the heterogeneous variants of BERT (22), ELECTRA (23), and RoBERTa (24), only the RoBERTa-based model shows higher performance than the competitors.

TABLE IV. THE PERFORMANCE OF DIFFERENT LABELS ON THE DEVELOPMENT DATASET.

Model	*Labels*	*Precision(%)*	*Recall(%)*	*F1(%)*	*Support Number*
BioBERT-Base	Treatment	86.91	91.12	88.96	2207
	Diagnosis	84.99	87.90	86.42	1546
	Prevention	**94.01**	**93.56**	**93.79**	2750
	Mechanism	89.51	81.08	85.09	1073
	Transmission	74.11	57.03	64.46	256
	Epidemic Forecasting	65.98	82.81	73.44	192
	Case Report	93.32	81.12	86.79	482

TABLE V. THE COMPARISON OF ALL SINGLE PRE-TRAINED MODELS.

Architecture	*EBP(%)*	*EBR(%)*	*EBF(%)*
BioBERT-Base	91.02	90.87	90.94
BioBERT-Large	90.34	90.23	90.28
PubMedBERT	92.02	**93.20**	92.61
CovidBERT	91.30	92.68	91.98
BioELECTRA	91.68	93.09	92.38
BioM-ELECTRA	**92.32**	92.60	92.46
BioMed_RoBERTa	92.19	93.12	**92.65**

C. Ensemble Learning Performance on The Test Dataset

In the official evaluation phase of LitCovid, each participant was allowed to submit up to five different predictions, and all the submissions would be evaluated using both label-based and instance-based metrics described before. For a fair comparison, the organizers also implemented a shallow embedding method of ML-Net (26), which was regarded as the official baseline system.

During the competition, we finally submitted five prediction results according to different ensemble policies. In particular, our first submission *Sub_1* applied the fine-tuned single model of *BioMed_RoBERTa* (21) for the challenge, which performs best among all the single pre-trained models on the development dataset.

Since there were seven different pre-trained models initialized in our approach, when fine-tuning these pre-trained models on the development dataset, we reserved those checkpoints for every single model which suggested the best performance on the F-measures of MaF, MiF, and EBF, respectively. Finally, 21 different checkpoints were totally reserved in which each pre-trained model possessed 3 checkpoints of its own. Our second submission *Sub_2* ensembled seven separate checkpoints that performed best on the measurement of MiF upon the development dataset. Correspondingly, the submissions of *Sub_3* and *Sub_4* ensembled the checkpoints with the best performance of MaF and EBF on the development dataset, respectively. Moreover, the *Sub_5* stands for the ensemble policy aggregating all 21 checkpoints of the fine-tuned pre-trained models. In the following, a comprehensive comparison between the official baseline system and our ensemble learning approach on the test dataset is performed.

Table VI reports the official statistics of the measures for all the submissions by participating systems, as well as the system performance of our proposed ensemble learning approach. In total, 80 valid submissions are considered for the statistics calculation. The detailed information of the mean, std, Q1, median, and Q3 of the submissions are listed in the table. It could be observed that the official baseline system of ML-Net (26) reaches decent achievements with the MaF of 76.55%, MiF of 84.37%, and EBF of 86.78%, respectively. Compared with the official baseline, the mean and median F-measurements of all submitted systems achieve more promising results. Additionally, although most submissions would outperform the baseline system, there are still relatively large standard deviations among all submissions.

In terms of our submission, the performance of *Sub_1* significantly outperforms the official baseline system as well as the median system. Excluding the submission of *Sub_1*, which merely utilizes the single pre-trained model of *BioMed_RoBERTa* (21), all the other submissions consistently achieve superior performance than the upper quartiles of Q3, carrying slight differences among each other. This indicates the effectiveness of the proposed ensemble learning policies. Furthermore, since the submission of *Sub_5* aggregating all fine-tuned checkpoints, it is able to further increase the final performance up to the MiF score of 91.39% and to the EBF score of 93.21, respectively. In a word, the experimental results indicate the effectiveness of the ensemble learning approach, due to its efficient aggregation of the multiple classifiers.

D. Error Analysis

To explore the challenging issues in practice and provide insights for future work, we closely analyzed the errors and grouped the main reasons as follows:

- ***Insufficient textual contents:*** As pre-trained models have some constraints about the length of the input text, some overlong texts need to be truncated and the essential information would be lost. This brings unexpected difficulties to the ensemble learning model during prediction.
- ***Complexity of language expression:*** If the target topic is not clearly expressed in the biomedical article, it is difficult for our ensemble learning approach to recommend labels.
- ***Prediction bias***: Since the topic labels follow an extremely imbalanced distribution, the prediction of our model is prone to pay more attention to the head labels aggressively while miss the tail ones conservatively.

TABLE VI. THE SYSTEM PERFORMANCE ON THE TEST DATASET.

Team submissions stats	***MaP(%)***	***MaR(%)***	***MaF(%)***	***MiP(%)***	***MiR(%)***	***MiF(%)***	***EBP(%)***	***EBR(%)***	***EBF(%)***
Mean	86.70	80.12	81.91	89.67	86.24	87.78	89.85	88.87	89.31
Std	6.09	7.94	7.01	5.41	4.82	4.29	5.21	4.51	4.60
Q1	84.63	75.45	76.51	88.03	84.52	85.41	86.99	86.19	86.68
Median	88.35	83.83	85.27	91.08	88.43	89.25	91.88	90.97	91.32
Q3	90.79	85.55	86.70	92.51	89.64	90.83	93.53	91.92	92.54
Baseline (ML-Net)	83.64	73.09	76.55	87.56	81.42	84.37	88.49	85.14	86.78
Sub_1 (Single Model)	88.40	84.00	85.51	89.91	88.74	89.32	91.70	91.20	91.45
Sub_2 (Ensemble by Best MiF)	**91.39**	85.34	**87.49**	92.12	**90.57**	91.34	93.53	92.79	93.16
Sub_3 (Ensemble by Best MaF)	90.16	**86.07**	87.42	92.17	90.49	91.32	93.55	**92.81**	93.18
Sub_4 (Ensemble by Best EBF)	90.78	84.85	86.92	**92.79**	89.99	91.37	**93.96**	92.43	93.19
Sub_5 (Ensemble All)	90.99	85.22	87.26	92.52	90.29	**91.39**	93.78	92.64	**93.21**

CONCLUSIONS AND FUTURE WORK

This research proposed an ensemble learning method for the COVID-19 multilabel classification problem, which utilized multiple biomedical pre-trained models. Particularly, it leveraged seven advanced models for initialization with homogeneous and heterogenous structures through an ensemble bagging manner. To enhance the representation abilities, it further employed additional biomedical knowledge to facilitate ensemble learning. The experimental results on the LitCovid datasets show the effectiveness of the proposed ensemble learning method.

Our research on ensemble learning exhibits promising results for the COVID-19 multilabel classification research on biomedical literature. In future work, we plan to develop more advanced deep learning algorithms with richer representation capabilities and introduce more sophisticated ensemble techniques to the current structure for better generalization.

ACKNOWLEDGMENT

This work is supported by the research grants of The Hong Kong Polytechnic University Projects #G-YW4H and #1-W182.

REFERENCES

1. Wang,L.L., Lo,K., Chandrasekhar, Y., et al. (2020) CORD-19: The Covid-19 Open Research Dataset. arXiv:2004.10706.
2. Esteva,A., Anuprit,K., Romain,P., et al. (2020) Co-search: Covid-19 information retrieval with semantic search, question answering, and abstractive summarization. arXiv:2006.09595.
3. Chen,Q., Allot,A., Lu,Z. (2021) LitCovid: an open database of COVID-19 literature. *Nucleic Acids Research*, 49(D1), D1534-D1540.
4. Chen,Q., Allot,A., Lu,Z. (2020) Keep up with the latest coronavirus research. *Nature*, 579(7798):193.
5. Betsch,C. (2020) How behavioural science data helps mitigate the COVID-19 crisis. *Nature Human Behaviour*, 4(5), 438-438.
6. Madabhavi,I., Sarkar,M., Kadakol,N. (2020) COVID-19: a review. *Monaldi Archives for Chest Disease*, 90(2).
7. Esakandari,H., Mohsen,N.A., Javad,F.A., et al. (2020) A comprehensive review of COVID-19 characteristics. *Biological Procedures Online*, 22,1-10.
8. Nentidis,A., Krithara,A., Bougiatiotis,K., et al. (2020) Overview of BioASQ 2020: The Eighth BioASQ Challenge on Large-Scale Biomedical Semantic Indexing and Question Answering. *In International Conference of the Cross-Language Evaluation Forum for European Languages*, 194-214.
9. Xun,G., Jha,K., Zhang,A. (2020) MeSHProbeNet-P: Improving Large-scale MeSH Indexing with Personalizable MeSH Probes. *ACM Transactions on Knowledge Discovery from Data*, 1-14.
10. Dai,S., You,R., Lu,Z., et al. (2020) FullMeSH: improving large-scale MeSH indexing with full text. *Bioinformatics*, 36(5),1533-1541.
11. Liu,K., Peng,S., Wu,J., Zhai,C., et al. (2015) MeSHLabeler: improving the accuracy of large-scale MeSH indexing by integrating diverse evidence. *Bioinformatics*, 31(12),i339-i347.
12. Chen,Q., Allot,A., Leaman,R., et al. (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.
13. Sagi,O. and Rokach,L. (2018) Ensemble learning: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4), e1249.
14. Wu,S., Wang,X., Wang,L. et al. (2020) Tencent Neural Machine Translation Systems for the WMT20 News Translation Task. Proceedings of the Fifth Conference on Machine Translation, 313-319.
15. Wang,X., Tu,Z. Shi,S. et al. (2020) Tencent AI Lab Machine Translation Systems for the WMT20 Biomedical Translation Task. Proceedings of the Fifth Conference on Machine Translation, 881-886.
16. Lee,J., Yoon,W., Kim,S., et al. (2020) BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4), 1234-1240.
17. Gu,Y., Tinn,R., Cheng,H., et al. (2020). Domain-specific language model pretraining for biomedical natural language processing. arXiv preprint arXiv:2007.15779.
18. Hebbar,S. and Xie,Y. (2021) CovidBERT-Biomedical Relation Extraction for Covid-19. *In The International FLAIRS Conference Proceedings*, 34.
19. Kanakarajan,K., Kundumani,B., Sankarasubbu,M. (2021) BioELECTRA: Pretrained Biomedical text Encoder using Discriminators. *In Proceedings of the 20th Workshop on Biomedical Language Processing*, 143-154.
20. Alrowili,S. and Vijay-Shanker,K. (2021) BioM-Transformers: Building Large Biomedical Language Models with BERT, ALBERT and ELECTRA. *In Proceedings of the 20th Workshop on Biomedical Language Processing*, 221-227.

21. Gururangan,S., Marasović,A., Swayamdipta,S., et al. (2020) Don't stop pretraining: adapt language models to domains and tasks. arXiv preprint arXiv:2004.10964.
22. 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.
23. 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.
24. Liu,Y., Ott,M., Goyal,N., et al. (2019) Roberta: A robustly optimized bert pretraining approach. arXiv:1907.11692.
25. Loshchilov,I. and Hutter,F. (2017) Decoupled weight decay regularization. arXiv:1711.05101.
26. Du,J., Chen,Q., Peng,Y., Xiang,Y., et al. (2019) ML-Net: multi-label classification of biomedical texts with deep neural networks. *Journal of the American Medical Informatics Association*, 26(11),1279-1285.