KnowLab at BioCreative VII Track 5 LitCovid: Ensemble of deep learning models from diverse sources for COVID-19 literature classification

Hang Dong[1,4], Minhong Wang[3], Huayu Zhang[2], Arlene Casey[2], Honghan Wu[2,3,4]
[1]Centre for Medical Informatics, Usher Institute, University of Edinburgh, Edinburgh, United Kingdom
[2]Advanced Care Research Centre, Usher Institute, University of Edinburgh, Edinburgh, United Kingdom
[3]Institute of Health Informatics, University College London, London, United Kingdom
[4]Health Data Research UK, London, United Kingdom

***Abstract*—Classifying scientific literature into an abstract set of topics requires leveraging various sources from the publication and external knowledge. In the BioCreative VII LitCovid track on COVID-19 literature multi-label topic annotation, we applied state-of-the-art deep learning based document classification models (BERT, variations of HAN, CNN, LSTM) and each with a different combination of metadata (title, abstract, keywords, and journal), knowledge sources, pre-trained embedding, and data augmentation techniques. Several ensemble techniques were then used to combine individual model outputs for synergized predictions. We showed that a class-specific average ensembling of the pre-trained and task-specific models achieved the best micro-F_1 score in validation (90.31%) and testing (89.32%) sets in the experiments, beyond the medium (89.25%) and mean value (87.78%) of all 80 valid submissions. We summarize lessons learned from our work on this task.**

Keywords—deep learning; ensemble learning; multi-label classification; document classification

I. Introduction

COVID-19 literature is growing rapidly with over 10,000 publications each month since spring 2020 (1). This large number of publications needs to be well curated for easy access by researchers and decision makers. The LitCovid project supports the curation of the literature and categorizes them into one of eight topics for easy browsing (1). This requires an automated annotation of the vast amount of growing literature.

LitCovid track at BioCreative VII is therefore organized as a shared task for COVID-19 literature topic annotation (2-3). We experimented with various deep learning based methods leveraging different metadata, models, knowledge sources, embedding, and data augmentation techniques. We show that a class-specific average ensembling of pre-trained and task-specific models produced the best results. We describe ours methods, results, and lessons learned in the following sections.

II. Methods

We formalize the topic annotation task as a multi-label classification problem. Deep learning has been well adapted for multi-label classification (4-8). One of the earliest work is since 2006, showing the advantages of neural networks compared to other methods (4). Deep learning based multi-label classification has been further applied to text classification in general (5) and for scientific literature and social texts (6), clinical notes (7-8), with superior results than conventional methods, e.g. Support Vector Machine (4, 6-7).

Table I on the next page summarizes the deep learning models and their key characteristics we applied for the shared task on scientific (COVID-19) literature classification. Our team organized the participation as an internal hackathon and each of our five members aimed to explore a distinct set of approaches for scientific literature classification. The intuition is that, the more *diverse* the individual approaches, the better the overall results that could potentially be achieved with a good model ensembling strategy (9). We describe the characteristics of the individual methods below.

A. Problem Formulation

Deep learning aims to learn a document representation v and approximates it to the multi-hot representation y of the labels where $y = [y_1, y_2, \dots, y_n]$ and y_i is a binary value (0 or 1) indicating the relevance of the label to the document (5). The *multi-label* classification model usually has a feedforward layer as its last layer which projects the document representation v to the logits s in the n-dimensional label space ($s \in R_n$). A logistic sigmoid activation then casts the logits s into a multi-hot prediction y'. The binary cross-entropy is used as the loss function which quantifies the distance between y and y' in a continuous manner. Alternatively, a common transformation of the multi-label problem is to treat it as n single *multi-class* classification problems. We denote the former multi-label formulation as ML and the multi-class formulation as MC. We follow the mainstream approach in the deep learning literature and used the ML formulation (5-8) for most of our methods.

B. Data Preprocessing and Representation

We focus on four aspects to pre-process and represent data.

Metadata Used The title and abstract information are the most relevant sources for a human to annotate a publication. Most of our models used both sources. Keywords are a direct categorization of a publication, thus relevant to the topic annotations. Journal also indicates the discipline of the work.

TABLE I. SCIENTIFIC LITERATURE CLASSIFICATION METHODS AND THEIR CHARACTERISTICS APPLIED TO THE BIOCREATIVE VII LITCOVID TRACK

Method ID	Method Name	Metadata Used	Model	Problem Formulation	Data Representation	Vocabulary or Knowledge Used	Data Augmentation
1	BlueBERT	Title + abstract	BlueBERT + FFNN	ML	BlueBERT (fine-tuning)	-	-
2	PubMedBERT-MLP	Title + abstract	MLP (with ReLU)	ML	PubMedBERT (as features)	-	-
3-4	JMAN and JMAN-BT	Title + abstract	JMAN	ML	CBOW from MIMIC-III discharge summaries (100dim)	-	w/ or w/o Back Translation (to German)
5-6	HLAN and HLAN-BT	Title + abstract	HLAN	ML		-	
7-8	HAGRU and HAGRU-BT	Title + abstract	HAGRU	ML		-	
9-10	HAN and HAN-BT	Title + abstract	HAN	ML		-	
11	CNN	Title + Abstract + Keywords	Multi-channel CNN (with ReLU)	ML	GloVe from Wikipedia 2014 + Gigaword 5 (100dim)	-	-
12	SJR-UMLS-MeSH-MLP	Journal + Title + Abstract	MLP (with ReLU)	ML	Bag of Words (921dim)	Journal categories (SJR) + UMLS (MedCAT) + MeSH (E-utilities)	-
13	UMLS-Bi-LSTM	Title + Abstract	Bi-LSTM + FFNN	MC	Sequence of UMLS concepts	UMLS (SemEHR)	-

Vocabulary Used We extracted concepts or vocabularies from the title and abstract. We used MedCAT (10) and SemEHR (11) to extract the Unified Medical Language System (UMLS) concepts and used EFetch in E-utilities (12) to query concepts in Medical Subject Headings (MeSH) from databases. We queried the data from Scimago Journal & Country Rank (SJR) (13) to obtain the disciplinary categories of each journal.

Data Representation Word embedding (e.g. Continuous Bag of Words (CBOW) (14) or GloVe (15)) and contextual embedding (e.g. BERT (16)) were used to represent the title and the abstract. We applied domain-specific language models, BlueBERT (17) and PubMedBERT (18), which have been pre-trained from texts in PubMed. For the concept annotations and the journal categories, Bag of Words was mainly applied.

Data Augmentation A key characteristic of multi-labelled data is class imbalance (19). Over-sampling with data augmentation is a key method to alleviate the issue of low frequent labels in imbalanced data. We used back translation (BT) for data augmentation, which automatically paraphrases a document after it being translated to another language (e.g. German) and back to English. The document-level semantics is not changed and the new document can share the same set of labels as the original document. We doubled the samples of the training documents for the n_a lowest frequent labels (n_a=3).

C. Models

The applied deep learning models include the large pre-trained, self-supervised language models, e.g. BERT (16), and the task-specific models, e.g. HAN (20). The pre-trained language models like BERT significantly outperforms task-specific models in a variety of tasks (16-17). But this is not always the case in the clinical and biomedical domain, e.g. multi-label clinical coding (8, 21).

For *pre-trained language models*, we fine-tuned BlueBERT with a feedforward neural network layer (FFNN) (method 1) and used the second-last layer of PubMedBERT as features with a multi-layer perceptron (MLP) (method 2).

For *task-specific models*, we applied Hierarchical Attention Network (HAN) (20) and its several variations, Joint Multi-label Attention Network (JMAN) (6), Hierarchical Label-wise Attention Network (HLAN) (8), and Hierarchical Attention Gated Recurrent Unit (HAGRU) (22). HAN uses word-level and sentence-level attention mechanisms to learn to select the specific parts of a document for topic annotation. JMAN encodes each title and abstract separately, and models the attention on each sentence (in the abstract) that is guided by the title. HLAN and HAGRU models further have label-wise attention mechanisms which generate a distinct document representation for each label. HAGRU does not apply the label-wise word-level attention mechanism compared to HLAN. We also adapted the Convolutional Neural Networks (CNN) (23) with multiple channels (method 11), MLP with Rectified Leaky Unit (ReLU) activation (method 12), and Bidirectional Long Short Term Memory (Bi-LSTM) (method 13) models for the task.

D. Model ensemble strategy

Model ensembling aims to derive a model of better performance by aggregating the results from different models. Ensemble learning has been applied to predict poor prognosis

in COVID-19 with a *data-specific* strategy, e.g. choosing the most competent model for each patient (24). Instead, we applied a *class-specific* ensemble of the models (choosing the best models for each label). For each label, we selected the top-k models with the best label-specific F_1 score and then averaged the raw, continuous-valued predictions of each model to produce the final model. This simple averaging can help produce a model with lower variance compared to the individual models (9). We discovered that k as 5 produced the best validation results in our 13 models listed in Table I.

We also experimented with other strategies on the validation set: a class-specific *naïve* ensembling approach that uses the best model for each class in terms of F_1 (i.e. k=1); or *class-agnostic* averaging, i.e. averaging the top-k predictions of best micro-F_1 scores of all labels; a class-agnostic averaging all of our models (i.e. k=13); and majority voting instead of averaging. Among them, the class-specific model averaging of top-5 models achieved the best micro-F_1 on the validation set.

III. Experiments

The overview of the BioCreative VII LitCovid track is in (3). We used the training, validation, and testing datasets from BioCreative VII LitCovid track on COVID-19 literature classification. There are 7 classes, sorted from the highest frequency to the lowest in the training set (with rounded frequencies in thousands): Prevention (11.1K), Treatment (8.7K), Diagnosis (6.2K), Mechanism (4.4K), Case Report (2.1K), Transmission (1.1K), Epidemic Forecasting (0.6K). In total, there are 24,960, 6,239, and 2,500 publication entries in the training, validation, and testing set, respectively. The PMID (PubMed ID), journal name (abbreviated), title, abstract, keywords, publication types, authors, and DOI were provided for each entry. We extracted UMLS, MeSH, and SJR vocabularies from the texts and journal names (see Vocabulary Used in Section B). For SJR-UMLS-MeSH-MLP (method 12), after filtering low frequent annotations, the final number of SJR, MeSH, UMLS concepts, and UMLS semantic types were 70, 139, 597, 115, respectively, altogether 921 dimensions.

We developed all of our models on the training set only and used the validation set for internal benchmarking and parameter tuning. We used the Microsoft Translator[1] to back translate the training documents (titles and abstracts) which are associated with the 3 classes of lowest frequencies, i.e. Case Report, Transmission, and Epidemic Forecasting. This further provided us 3,736 documents for training.

Our models were implemented using Tensorflow (version 1 or 2) or PyTorch (for fine-tuning BlueBERT with Huggingface Transformers (25)). We also used BERT-as-service[2] to extract the second-last layer of PubMedBERT. The maximum length for BERT models were 512, beyond 99.47% of training documents. The hidden sizes of the 4-layer MLP for method 2 and 12 were [32,32,32,16]. The hidden size for the variations of HAN (methods 3-10) were 100. For CNN, all three channels were with static embedding; the kernel sizes were [4,8,12] and the number of filters was 32. The dropout rate for HAN variations and CNN was 0.5. In method 13, the Bi-LSTM had 2 layers and FFNN had 3 layers.

[1] https://docs.microsoft.com/en-us/azure/cognitive-services/translator/

[2] https://github.com/hanxiao/bert-as-service

TABLE II. Micro-level Testing Results of Submitted Models

	Precision	Recall	F_1
Mean	0.8967	0.8624	0.8778
Q1	0.8803	0.8452	0.8541
Median	0.9108	0.8843	0.8925
Q3	**0.9251**	**0.8964**	**0.9083**
Individual model			
BlueBERT	0.8986	**0.8850**	0.8917
Class-agnostic averaging			
top-k (k=5)	0.9184	0.8626	0.8896
all (k=13)	**0.9198**	0.8501	0.8836
Class-specific averaging			
naïve (k=1)	0.9183	0.8637	0.8901
top-k (k=5)	0.9165	0.8711	**0.8932**

TABLE III. Micro-Level Validation Results of Selected Models

	Precision	Recall	F_1
BlueBERT	0.8828	**0.8990**	**0.8908**
PubMedBERT-MLP	**0.8980**	0.8673	0.8824
JMAN	0.8927	0.8668	0.8796
JMAN-BT	0.8725	0.8683	0.8704
CNN	0.8686	0.7662	0.8142
SJR-UMLS-MeSH-MLP	0.7364	0.7344	0.7354
UMLS-Bi-LSTM	0.8080	0.8160	0.8120
Class-specific ave. top-5	**0.9128**	0.8936	**0.9031**

IV. Results

Our testing results in the LitCovid shared task are presented in Table II. Our best performance (89.32% micro-F_1 score) was achieved by the class-specific averaging of the top-5 models (two BERT models and three HAN variations), better than the median, mean and lower quartile (Q1) of all 80 valid submissions to the shared task. Our result is below the upper quartile (Q3), which may be partly because our team did not include the validation set within the final training of models.

Ensemble of models improved micro-level precision and F_1 over single models, with a slight drop of recall, compared to our best performing single model, BlueBERT (Table II-III). We can also observe in the testing results (Table II) that ensembling the best models for each class (class-specific) performed better than simply ensembling the models of best micro-F_1 regardless of classes (class-agnostic). This is because for certain low-frequent labels, e.g. Transmission, models trained using augmented data with back translation performed significantly better (e.g. improved from 57.8% to 62.5% with JMAN for Transmission), but this was not the case for high-frequent labels, e.g. Treatment.

Table III also shows the validation results of selected models. The two BERT models (fine-tuning or as features) performed significantly better than task-specific models (JMAN and CNN). Titles and abstracts were the most relevant sources for topic annotation, compared to concept annotations (UMLS and MeSH), keywords, and journal categories.

V. Conclusions and Lessons Learned

In this paper, we described our deep learning models and their ensembling strategies for multi-label topic annotation of publications in the BioCreative VII LitCovid shared task. Our experimental results showed that a class-specific averaging of both pre-trained language models and task-specific models perform the best in terms of micro-level F_1.

We summarize our lessons learned based on the results above, which will empower our future natural language processing applications in the clinical and biomedical domain:

Embrace pre-trained language models. The significantly better performance of BERT suggests us focusing on self-supervised contextual embedding for document representation.

Ensemble pre-trained language models with task-specific models. Simply averaging the raw predictions of BERT models and variations of HAN achieved better micro-F_1. More advanced ensemble learning needs to be explored.

Tackle challenging labels with data augmentation and domain knowledge. In this task, Transmission and Epidemic Forecasting were the two most challenging labels for all models. Data augmentation strategies were shown effective. Back translation significantly improved the performance of classifying the Transmission class. Detailed data analysis may further help understand the difficult labels. We suggest that a domain expert focuses on the two difficult labels to identify rules and ontology concepts to complement the BERT models.

Acknowledgment

Our 2-day hackathon was supported by Usher Institute Small Grant. We also appreciate the GPU server provided by the Language Technology Group, University of Edinburgh.

References

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

2. BioCreative VII. Track 5 - LitCovid track Multi-label topic classification for COVID-19 literature annotation. Available from: https://biocreative.bioinformatics.udel.edu/tasks/biocreative-vii/track-5/

3. Chen, Q., Allot, A., Leaman, R., Doğan, R. I., Lu, Z. (2021, November). 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. Zhang, M. L., & Zhou, Z. H. (2006). Multilabel neural networks with applications to functional genomics and text categorization. *IEEE transactions on Knowledge and Data Engineering*, **18**(10), 1338-1351.

5. Nam, J., Kim, J., Mencía, E. L., Gurevych, I., & Fürnkranz, J. (2014, September). Large-scale multi-label text classification—revisiting neural networks. In *Joint european conference on machine learning and knowledge discovery in databases* (pp. 437-452). Springer, Berlin, Heidelberg.

6. Dong, H., Wang, W., Huang, K., & Coenen, F. (2020). Automated social text annotation with joint multilabel attention networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(5), 2224-2238.

7. Karimi, S., Dai, X., Hassanzadeh, H., & Nguyen, A. (2017, August). Automatic diagnosis coding of radiology reports: a comparison of deep learning and conventional classification methods. In *BioNLP 2017* (pp. 328-332).

8. Dong, H., Suárez-Paniagua, V., Whiteley, W., & Wu, H. (2021). Explainable automated coding of clinical notes using hierarchical label-wise attention networks and label embedding initialisation. *J. Biomed. Inf.,* **116**, 103728.

9. Polikar, R. (2012). Ensemble Learning. In C. Zhang & Y. Ma (Eds.), *Ensemble Machine Learning: Methods and Applications* (pp. 1–34). Springer US.

10. Kraljevic, Z., Searle, T., Shek, A., Roguski, L., Noor, K., Bean, D., ... & Dobson, R. J. (2021). Multi-domain clinical natural language processing with medcat: the medical concept annotation toolkit. *Artificial Intelligence in Medicine*, **117**, 102083.

11. Wu, H., Toti, G., Morley, K. I., Ibrahim, Z. M., Folarin, A., Jackson, R., ... & Dobson, R. J. (2018). SemEHR: A general-purpose semantic search system to surface semantic data from clinical notes for tailored care, trial recruitment, and clinical research. *Journal of the American Medical Informatics Association*, **25**(5), 530-537.

12. Sayers, E. (2010). *A General Introduction to the E-utilities. Entrez Programming Utilities Help* [Internet]. Bethesda (MD): National Center for Biotechnology Information (US). Available from: https://www.ncbi.nlm.nih.gov/books/NBK25497/

13. Journal Rankings (2020). Scimago Journal & Country Rank. Available from: https://www.scimagojr.com/journalrank.php?out=xls

14. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

15. Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In *EMNLP 2014* (pp. 1532-1543).

16. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *NAACL-HLT 2019* (pp. 4171-4186).

17. Peng, Y., Yan, S., & Lu, Z. (2019, August). Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. In *BioNLP 2019* (pp. 58-65).

18. Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., ... & Poon, H. (2020). Domain-specific language model pretraining for biomedical natural language processing. arXiv preprint arXiv:2007.15779.

19. Gibaja, E., & Ventura, S. (2015). A tutorial on multilabel learning. *ACM Computing Surveys (CSUR)*, **47**(3), 1-38.

20. Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., & Hovy, E. (2016, June). Hierarchical attention networks for document classification. In *NAACL-HLT 2016* (pp. 1480-1489).

21. Ji, S., Hölttä, M., & Marttinen, P. (2021). Does the Magic of BERT Apply to Medical Code Assignment? A Quantitative Study. arXiv preprint arXiv:2103.06511.

22. Baumel, T., Nassour-Kassis, J., Cohen, R., Elhadad, M., & Elhadad, N. (2018, June). Multi-label classification of patient notes: case study on ICD code assignment. In *Workshops at the thirty-second AAAI conference on artificial intelligence, Health Intelligence*.

23. Kim, Y. (2014, October). Convolutional Neural Networks for Sentence Classification. In *EMNLP 2014* (pp. 1746-1751).

24. Wu, H., Zhang, H., Karwath, A., Ibrahim, Z., Shi, T., Zhang, X., ... & Guthrie, B. (2021). Ensemble learning for poor prognosis predictions: A case study on SARS-CoV-2. *Journal of the American Medical Informatics Association*, **28**(4), 791-800.

25. Wolf, T., Chaumond, J., Debut, L., Sanh, V., Delangue, C., Moi, A., ... & Rush, A. M. (2020, October). Transformers: State-of-the-art natural language processing. In *EMNLP 2020* (pp. 38-45).