# Bio-Medical Multi-label Scientific Literature Classification using LWAN and Dual-attention module

**Deepanshu Khanna[$], Aakash Bhatnagar[%], Nidhir Bhavsar[&], Muskaan Singh[#]** and **Petr Motlicek[#]**
[$] ECED, Thapar Institute of Engineering and Technology, India
[%] Boston University, Boston, Massachusetts
[&]University of Potsdam, Potsdam, Germany
[#] Speech and Audio Processing Group, IDIAP Research Institute, Martigny, Switzerland
dkhanna_be19@thapar.edu, aakash07@bu.edu,
bhavsar@uni-potsdam.de, (msingh,petr.motlicek)@idiap.ch

## Abstract

An enormous amount of research has been undertaken to overcome the severe impact of COVID-19 pandemic. These scientific findings are being reported in biomedical literature at a significant rate of 10,000 articles/month. In this paper, we tackle automated topic annotation for COVID-19 literature using SPECTER, Bioformer, and PubMedBERT embeddings using Label-Wise Attention Network (LWAN) based Multi-Label Document Classification (MLDC) using Dual-attention module. We also include literature from cardiovascular domain, to generalise our proposed approach. We significantly, achieve 87.71%,72.83% and 79.75% F1-score on LitCovid, Obsumed, WHO-Covid datasets. We release our code-base here https://github.com/Deepanshu-beep/MLDC_LWAN_Attention.

## 1 Introduction

COVID-19 pandemic, has caused various unexpected challenges to public health and similar line of work. Since its outbreak, there has been a drastic loss in human life, leading to the exponential growth of research and innovation in this field , nearly 20000 articles ( till December 2019) have been published.

Due to the information overload, it became a tremendous task for the general public and research professionals to keep up the pace with the latest COVID-19 research. The rate of increase in publications related to the pandemic still continues to increase rapidly today.

Most of the biomedical literature focuses on multiple topics such as treatment, diagnosis, prevention, vaccine etc. These literature are often classified under multiple labels and therefore presents, Multi-Label Document Classification (MLDC) problem at a large scale.

Previously, various researchers have performed thorough experimentation over MLDC. The task of MLDC has been covered in various applications especially in medical domain. One of the most relevant datasets for the medical domain is the MIMIC-III [1], which contains an extensive literature of clinical notes for 16 ICD-9 codes. Similarly, LitCovid is another important dataset that can be used for MLDC, which contains various articles related to COVID-19 corresponding to their 7 unique labels.

Traditional approaches for MLDC included extraction of handcrafted features from documents and then using single or multiple classifiers as in [2]. Earlier works in this domain used Convolutional Neural Network (CNN) [3] and Seq-2-Seq [4]. Authors in [3], later extended their work in [5] by using a 1-dimensional convolutional network to learn text representations and evaluating their approach over 6 datasets. Another work [6] proved the effectiveness of systems involving attention mechanisms by combining Recurrent Neural Networks (RNNs) and self-attention network for MLDC. The paper [6] was one of the few works carrying out a comparison between probabilistic label trees and neural models. Recent top-performing models include methodologies such as Transfer learning, Few and Zero-Shot learning, and Label-Wise Attention Networks (LWANs). Articles [7, 8] experimented with transfer learning

using pre-trained language models such as BERT and ELMo, respectively. In [9], authors proposed a Zero-shot attention-based CNN network, which outperformed Zero-shot and Few Shot learning-based methods. Another prominent methodologies included combining LWANs along with BERT [10]. Their work compared LWAN with attention-based RNNs and Hierarchical Attention Network (HAN). Multiple variations of LWANs have been used widely for MLDC, such as CNN-LWAN [11], Z-CNN-LWAN [9]. Hence using LWAN for MLDC became a strong motivation for our work.

In this paper, (1) We tackle the issue of large scale MLDC especially for medical domain by classifying various articles related to COVID-19 and cardiovascular diseases to their classes. (2) We tackle automated topic annotation for COVID-19 and cardiovascular literature using SPECTER [12], Bioformer [13], and PubMedBERT [14] embeddings along with LWAN and Dual-attention module based architecture for Multi-Label Document Classification. (3) We evaluate the performance of the proposed architecture over two COVID-19 articles based databases: LitCovid [15] and WHO-Covid, and another Cardiovascular diseases based database: Ohsumed [16].

## 2 Methodology

We present the proposed system in Figure 1. Since it is crucial to preserve the contextual importance of the paper to predict its label, we decide to proceed with the classical transformer-based approach, unlike any other neural architecture. Evidently, the transformers-based approach performs eminently compared to CNN or LSTM network-based approaches, which have been discussed later.

Firstly, we create the embeddings for document representations, by concatenating the title and the abstract of the papers using the [CLS] and [SEP] tokens, represented as:

$$< [CLS], title, [SEP], abstract, [CLS] > \quad (1)$$

Secondly, we feed the above input sequence to experiment with three different word embedding methods namely, SPECTER, Bioformer, and PubMedBERT. As these three embedding models are pre-trained over biomedical literature, they were the most suited for this task.

- SPECTER is a pre-trained transformer based pre-trained language model which is an extension of SciBERT [17] and uses citation aware graphs. SPECTER creates high-quality document representations since, unlike other BERT-based models trained on intra-document information, SPECTER was trained on citations as inter-document information. Thus, providing high-quality document-level representations.

- Bioformer is another transformer language model pre-trained over PubMed abstracts and 1 million randomly sampled PubMed central full-text papers. Since LWANs make the architecture computationally expensive, we experimented with Bioformer being a lightweight model with nearly 60% fewer parameters than other BERT-based models. Moreover, it encodes biomedical text efficiently, and the input text length is 20% higher than the other BERT models.

- PubMedBert is also used for experimentation with the proposed network. This state-of-the-art model was pre-trained over PubMed abstracts. The reason behind its exceptional performance is that rather than continuing pre-training of other domain-specific language models over general-domain language models, it was pre-trained from scratch over biomedical data. Hence, behaving exceptionally in various NLP applications.

Finally, we utilise dual attention and LWAN network. A single self attention layer helps model to retain important information in an instance. To further improve this, we implied a dual-attention module that helps in generating relationship between different instances. Dual-attention lead to a significant improvement in results, specially for the classes that have fewer number of instances. In MLDC, LWAN is important because it helps in retaining class-wise information of every instance. Upon retrieving the document representations, we use a Dual-attention module comprising two sequential attention modules. The self-attention module takes word embeddings as input and generates contextualized word embeddings. This attention mechanism is performed by comparing each word with every word in a sentence and recomputing its weights according to the contextual

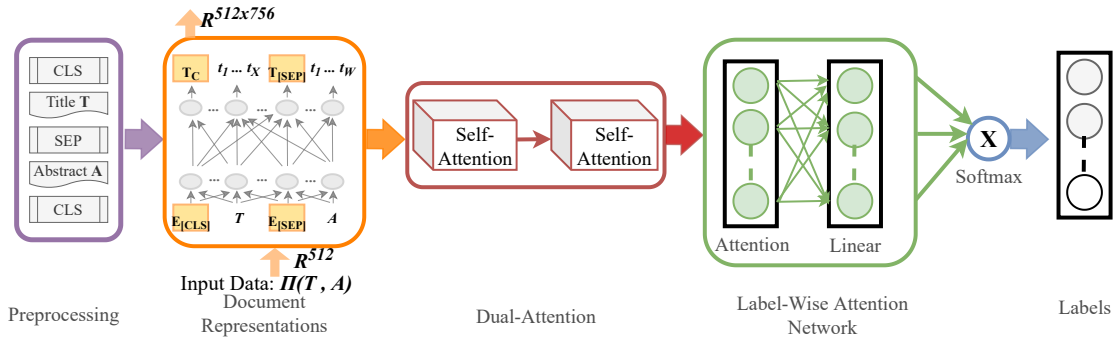


Figure 1: Proposed Methodology

relevance. The attention scores are calculated using three unique vectors: Key (K), Value (V), and Query (Q). The whole self-attention block can be divided in three sub-tasks:

- Calculating dot product similarity between the Query and Key vector $(Q \,.\, K_i)$. Upon computing the dot product, we determine the alignment scores that depict the semantic similarity of the pairs. Hence, we get to know the words that require higher attention scores.
- Further, these alignment scores are normalized using a non-linear activation function: Softmax, denoted by $a_i$.
- Finally, the total attention of words is calculated by multiplying the weights obtained $(a_i)$ and the original document representations embeddings $(V_i)$. The sum of these weighted scores is the output of the self-attention module as shown below:

$$\text{Attention}(Q, K, V) = \sum_i a_i V_i \tag{2}$$

Equation 2: Where: a is Attention score vector, V is the vector corresponding to the embeddings for document representation of Query (Q) and Key (K) vectors.

Using the double self-attention mechanism focuses more on the input information in a sentence and generates better attention vectors as compared to a single self-attention module. Since the self-attention mechanism generates embeddings with relationship amongst the input instances only, which is even more precise when using the Double-attention mechanism, it disregards the output completely. Therefore, to overcome this limitation, we introduce the use of Label-Wise-Attention-Network (LWAN) in our model that provides the attention for each label to classify in the dataset.

LWANs play a significant role in tasks related to multi-label prediction. They assign particular attention scores for every attention node corresponding to the output label while focusing on words with higher attention in input as well. Here, the attention scores are calculated by applying the same attention mechanism for the unique labels. The mathematical procedure for calculating attention scores for LWAN is also show below:

$$z_{i,l} = w_{a,l} h_i + b_{a,l} \tag{3}$$

$$\alpha_{i,l} = \frac{e^{z_{i,l}}}{\sum_{j=1}^{N} e^{z_{j,l}}} \tag{4}$$

$$s_l = \sum_{j=1}^{N} \alpha_{j,l} h_j \tag{5}$$

$$\beta_l = w_{f,l} s_l + b_{f,l} \tag{6}$$

$$p_l = \frac{e^{\beta_l}}{\sum_{r=1}^{L} e^{\beta_r}} \tag{7}$$

Equation 3-7: Where: $w_{a,l}$, $b_{a,l}$ correspond to weight and bias vector and $h_i$ is the hidden LSTM representation of the $i^{th}$ word. $\alpha_{i,l}$ are the attention weights obtained after normalizing $z_{i,l}$ using Softmax function. Further we obtain the whole sentence representation by computing the weighted average of $h_i$. Finally, we obtain the confidence score of $l_{th}$ label

| | LitCovid | Ohsumed | WHO-Covid |
|---|---|---|---|
| No. of unique abstracts | 31199 | 34389 | 169292 |
| No. of categories | 7 | 23 | 19 |
| Average sentences | 10 | 9 | 10 |
| Average tokens for input | 212 | 182 | 216 |
| Total no. of tokens | 6618763 | 6248274 | 36605199 |

Table 1: Statistics of LitCovid, Ohsumed and WHO-Covid databases.

| Model | F1-score |
|---|---|
| Team BJUT-BJFU @ Biocreative | 78.47 |
| KimCNN | 83.45 |
| LSTM | 83.95 |
| LSTM$_{reg}$ | 84.05 |
| XML-CNN | 84.2 |
| SPECTER-Dual Attention-LWAN | 87.13 |
| Bioformer-Dual Attention-LWAN | 87.64 |
| PubMedBERT-Dual Attention-LWAN | 87.71 |
| **Bioformer** | **88.75** |

Table 2: F1-score performance over LitCovid database for different embeddings used for document representations and comparison with various models.

| Model | F1-score |
|---|---|
| Sentence2Vec | 37.34 |
| LSTM | 62.7 |
| HAN | 67.0 |
| TextING | 69.5 |
| HAN+TextING | 70.3 |
| Bioformer-Dual Attention-LWAN | 71.17 |
| HGMETA | 72.0 |
| PubMedBERT-Dual Attention-LWAN | 72.48 |
| **SPECTER-Dual Attention-LWAN** | **72.83** |

Table 3: F1-score performance over Ohsumed database for different embeddings used for document representations.

| Model | F1-score |
|---|---|
| Bioformer-Dual Attention-LWAN | 75.61 |
| PubMedBERT-Dual Attention-LWAN | 78.87 |
| **SPECTER-Dual Attention-LWAN** | 79.75 |

Table 4: F1-score performance over WHO-Covid database for different embeddings used for document representations.

denoted by $\beta_l$, which is further normalized using Softmax function and denoted as: $p_l$.

## 3 Experimental Details

In this section, we present, our experimental settings, with dataset in section 3.1, hyperparameter in section 3.2 and training in section 3.3.

### 3.1 Datasets

We use three datasets: LitCovid, Ohsumed, and WHO-Covid, to experiment with our proposed architecture and prove its efficiency. A detailed description of the datasets has been given below. Various statistics of the datasets have also been depicted in the Table 1.

- LitCovid is a large corpus of articles published in PubMed about the COVID-19 and SARS-COV-2. The dataset contains in total 31199 unique articles corresponding to 7 classes, namely: Case Report, Diagnosis, Epidemic Forecasting, Mechanism, Prevention, Transmission and Treatment. We used 24960, and 6239 articles for our train and dev set, respectively.
- Ohsumed dataset is a subset of the MEDLINE database containing medical abstracts MeSH categories from 1991. The dataset is aimed to classify 23 cardiovascular disease categories. Ohsumed contains 34389 unique abstracts, out of which we used 27511 and 6878 for the training and validation.
- WHO-Covid Since the global pandemic of COVID-19, WHO has been collecting global literature on COVID-19. This database gets updated regularly from Monday to Friday. The database consists of multilingual content from searches of bibliographic databases, manually searched articles, and various experts referred scientific articles. During the time we scraped the articles, it had in total 169292 articles for the English language that were selected for experimentation.

### 3.2 Hyperparameter Setup

In our experiments, we analyze the performance of various embeddings for creating document representations. Table 2,3 and 4 show the performance of the proposed model with different embeddings experimented over the following databases.

As discussed earlier, one of the common challenges in MLDC is the imbalanced distribution of labels for training. Initially, we tackled the issue by adopting the weighted Binary-Cross entropy (BCE) loss function, which improved the performance by a slight margin. Later, we assign weight to each class to focus on minority labels as well rather than only on dominant labels. We calculate the weight for each class using the below formula:

$$W_i = \frac{C_m}{C_i} \tag{8}$$

Equation 8: Where: $C_i$ represents the count of the $i^{th}$ label, $C_m$ represents the count of the most dominant label and $W_i$ corresponds to the weight of the label.

### 3.3 Training

For the LitCovid database, we trained our model over the original train and dev set released, but for Ohsumed and WHO-Covid databases, we split the data in 80:20 for training and evaluation, respectively. We used the Weighted BCE loss function and trained for ten epochs for LitCovid and Ohsumed databases while for 25 epochs for the WHO-Covid database, due to its large size. We use a learning rate scheduler that changes the learning rate to 2 x $10^{-5}$ and linearly went down until 0. Other details of the hyperparameters have been shown in Table 5.

### 3.4 Result and Analysis

In our performance analysis, SPECTER performed best out of the 3 document representations selected. We achieved 87.71, 72.48, and 79.75 macro F1-score for LitCovid, Ohsumed, and WHO-Covid databases. Though the highest score we achieved for the LitCovid database is 87.71 using PubMedBERT embeddings, which is just 0.46% higher than that of SPECTER, hence proving its overall efficiency. Also, we observe that assigning class weights significantly improves the macro F1 scores for rare classes. For instance, we achieved an F1 score of 77.96 for the Epidemic Forecasting label in the LitCovid database, which is the rarest class in it.

The proposed architecture, despite its simple network, performs remarkably. For the LitCovid database, the current top-performing architecture [13] used the Bioformer language model fine-tuned over the database and achieved an F1 score of 88.75, 1.18% higher than the proposed model. We outperform various deep CNN based models: KimCNN [18], XML-CNN [3] along with regularized and unregularized LSTM [19]. Some of the examples predicted by the proposed model have been shown in Table 6. In [20], experimented with four different models: FastText, TextRCNN, TextCNN and Transformer. Their top-performing model achieved an F1-score of 78.47.

For the Ohsumed database, all of our experimented document representations outperform existing models. [21] presented HGMETA that initially extracts fusion embeddings of hierarchical semantics dependence and graph structure in a structured text. Further, they use a hierarchical LDA module and a structured text embedding module to merge the extracted hierarchical features with structured text information. HGMETA being the best performing model on Ohsumed yet, achieved an F1-score of 0.72. Another graph-based method, TextING [22] that treated each document as an individual graph while training, achieved an F1-score of 0.695. On the other hand, HAN [23] that is most commonly used method for structured text classification, uses a RNN-based network along with hierarchical attention mechanism achieved an F1-score of 0.67. Performances of various other baseline methods over Ohsumed database have been shown in Table 3.

## 4 Conclusion and Future Work

We present our proposed approach using Dual-Attention LWAN method and experiment with three different embeddings namely: SPECTER, Bioformer and PubMedBERT, for document representations. We evaluated the performance of the proposed architecture over three databases: two related to COVID-19 articles and the other one general biomedical articles based. We also observe that despite LWANs focusing just on input data, they perform remarkably when combined with a dual attention module. The only limitation faced during the implementation is the computational expense of LWAN when training upon databases with a large number of labels. As future work, we plan to mitigate this limitation by exploring other novel architectures and methodolo-

| Model | Hyperparameters | Number of parameters |
|---|---|---|
| SPECTER-Dual Attention-LWAN | learning rate: $2 \times 10^{-5}$<br>max sequence length: 512<br>batch size: 4 , 32<br>epochs: 10 , 25<br>model: SPECTER<br>warmup proportion: 0.2<br>dropout probability: 0.1 | model: $109M$<br>Attention: $1.8M$ |
| PubMedBERT-Dual Attention-LWAN | learning rate: $2 \times 10^{-5}$<br>max sequence length: 512<br>batch size: 4 , 32<br>epochs: 10 , 25<br>model: PubMedBERT (Abstract + Full-text)<br>warmup proportion: 0.2<br>dropout probability: 0.1 | model: $109M$<br>Attention: $1.8M$ |
| Bioformer-Dual Attention-LWAN | learning rate: $2 \times 10^{-5}$<br>max sequence length: 512<br>batch size: 4 , 32<br>epochs: 10 , 25<br>model: Bioformer-Cased<br>warmup proportion: 0.2<br>dropout probability: 0.1 | model: $42.5M$<br>Attention: $786K$ |

Table 5: Hyperparameter details for he experimental setup, for spectar-dual attention-LWAN, PuBMedBERT-Dual Attention-LWAN and Bioformer-Dual Attention

| **Article** | **Actual** | **Predicted** |
|---|---|---|
| Cardiac dysfunction in Multisystem Inflammatory Syndrome in Children | Clinical Practice Guide<br>Observational Study<br>Prognostic Study<br>Risk Factors | Observational Study<br>Risk Factors |
| I Told You the Invisible Can Kill You": Engaging Anthropology as a Response in the COVID-19 Outbreak in Italy". | Etiology Study<br>Observational Study | Etiology Study<br>Observational Study<br>Risk Factors |

Table 6: Examples of predictions by our model (Green color indicates the correct predictions, red color indicates incorrect predictions.

gies along with handling the data imbalance more effectively to create more effective MLDC models.

## 5 Acknowledgements

This work was supported by the European Union's Horizon 2020 research and innovation program under grant agreement No. 833635 (project ROXANNE: Real-time network, text, and speaker analytics for combating organized crime, 2019-2022).

## References

[1] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[2] Jinseok Nam, Jungi Kim, Eneldo Loza Mencía, Iryna Gurevych, and Johannes Fürnkranz. Large-scale multi-label text classification—revisiting neural networks. In *Joint european conference on machine learning*

*and knowledge discovery in databases*, pages 437–452. Springer, 2014.

[3] Jingzhou Liu, Wei-Cheng Chang, Yuexin Wu, and Yiming Yang. Deep learning for extreme multi-label text classification. In *Proceedings of the 40th international ACM SIGIR conference on research and development in information retrieval*, pages 115–124, 2017.

[4] Venkatesh Umaashankar and Girish Shanmugam S. Multi-label multi-class hierarchical classification using convolutional seq2seq. In *KONVENS*, 2019.

[5] Yang Liu, Hua Cheng, Russell Klopfer, Matthew R Gormley, and Thomas Schaaf. Effective convolutional attention network for multi-label clinical document classification. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5941–5953, 2021.

[6] Ronghui You, Zihan Zhang, Ziye Wang, Suyang Dai, Hiroshi Mamitsuka, and Shanfeng Zhu. Attentionxml: Label tree-based attention-aware deep model for high-performance extreme multi-label text classification. *Advances in Neural Information Processing Systems*, 32, 2019.

[7] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. Large-scale multi-label text classification on eu legislation. *arXiv preprint arXiv:1906.02192*, 2019.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[9] Anthony Rios and Ramakanth Kavuluru. EMR coding with semi-parametric multi-head matching networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2081–2091, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[10] Ilias Chalkidis, Manos Fergadiotis, Sotiris Kotitsas, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. An empirical study on large-scale multi-label text classification including few and zero-shot labels. *arXiv preprint arXiv:2010.01653*, 2020.

[11] James Mullenbach, Sarah Wiegreffe, Jon Duke, Jimeng Sun, and Jacob Eisenstein. Explainable prediction of medical codes from clinical text. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1101–1111, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[12] Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. SPECTER: document-level representation learning using citation-informed transformers. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 2270–2282. Association for Computational Linguistics, 2020.

[13] Li Fang and Kai Wang. Team bioformer at biocreative vii litcovid track: Multic-label topic classification for covid-19 literature with a compact bert model. In *Proceedings of the seventh BioCreative challenge evaluation workshop*, 2021.

[14] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23, 2021.

[15] Qingyu Chen, Alexis Allot, and Zhiyong Lu. Litcovid: an open database of covid-19 literature. *Nucleic acids research*, 49(D1):D1534–D1540, 2021.

[16] William Hersh, Chris Buckley, TJ Leone, and David Hickam. Ohsumed: An interactive retrieval evaluation and new large test collection for research. In *SIGIR'94*, pages 192–201. Springer, 1994.

[17] Iz Beltagy, Arman Cohan, and Kyle Lo. Scib-

ert: Pretrained contextualized embeddings for scientific text. *CoRR*, abs/1903.10676, 2019.

[18] Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, October 2014. Association for Computational Linguistics.

[19] Ashutosh Adhikari, Achyudh Ram, Raphael Tang, and Jimmy Lin. Rethinking complex neural network architectures for document classification. In *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)*, pages 4046–4051, 2019.

[20] Shuo Xu, Yuefu Zhang, and Xin An. Team bjut-bjfu at biocreative vii litcovid track: A deep learning based method for multi-label topic classification in covid-19 literature.

[21] Shaokang Wang, Li Pan, and Yu Wu. Meta-information fusion of hierarchical semantics dependency and graph structure for structured text classification. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 2022.

[22] Yufeng Zhang, Xueli Yu, Zeyu Cui, Shu Wu, Zhongzhen Wen, and Liang Wang. Every document owns its structure: Inductive text classification via graph neural networks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 334–339, Online, July 2020. Association for Computational Linguistics.

[23] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489, 2016.