LRL_NC at BioCreative VII LitCovid Track: Multi-Label Classification of COVID-19 Literature using ML-Based Approaches

Kushagri Tandon, Niladri Chatterjee
Department of Mathematics, Indian Institute of Technology, New Delhi, India

***Abstract*— The aim of the present paper is to propose techniques for automatic classification of documents pertaining to the COVID-19 global pandemic based on their topics. Here, the task of topic annotation is considered as a multi-label classification problem. The label set consists of seven different topics covering various aspects of the pandemic as given for the task 'LitCovid track multi-label topic classification for COVID-19 literature annotation'. Various machine learning based approaches have been tried to that effect. Experiments yielding the best results are discussed here. Two broad types of techniques are used, namely a partition based multi-label classification approach using tf-idf and log entropy representation as features; and a topic modelling based approach using document-topic distribution extracted from contextual topic models as features. The discussed partition-based methods performed best amongst the conducted experiments and have shown improvements upon the baseline performance with respect to most of the metrics used for evaluation.**

Keywords— Topic annotation; COVID-19; UMLS; Tf-idf; Log-entropy; Contextual Topic Modelling

I. Introduction

The objective of the present work is to construct a system for automatic topic annotation of COVID-19 literature. Since the emergence of COVID-19 as a global pandemic, the electronic world is full of documents providing different kinds of information, such as diagnosis, prevention, treatment, among others. Classification of these documents into some coherent groups/clusters will be highly beneficial for their use and utilization in future. However, the major challenge lies in the fact that a document's contents may not always be focusing on one single topic; rather it is likely to have contents which may be ascribed to more than one topic. This allows one to think of casting the aforesaid problem as a multi-label text categorization problem, which aims at assigning a subset of labels to a document or word sequence describing the instance in consideration. In particular, the present work reports experiments and results corresponding to the task, 'LitCovid track multi-label topic classification for COVID-19 literature annotation' [1] consisting of articles from LitCovid [2,3], which is a database for COVID-19-related papers from PubMed. The set of possible labels comprises seven topics, namely Treatment, Diagnosis, Prevention, Mechanism, Transmission, Epidemic Forecasting, and Case Report. Training, development, and test datasets consisting of 24,960, 6,239 and 2500 articles from LitCovid, respectively, have been provided for the task. Out of the fields provided for each article in the dataset, title, abstract, and keywords were used as the raw text input to create a system for automatic topic annotation.

In the present work two broad types of techniques for multi-label text categorization have been used. The first technique is a data-centric approach which uses insights on label co-occurrence patterns from the training data to segment the given multi-label classification problem into sub-problems, each of which is solved independently using the Random k-Labelsets [6] algorithm for multi-label classification. The experiments with this method are carried out in two parts: (i) on the raw text input, (ii) on the UMLS (Unified Medical Language System) [8] concepts extracted from this raw text input. On the other hand, the second technique uses document-topic distribution extracted from contextual topic models as features for a Binary Relevance multi-label classifier with LGBM (Light Gradient Boosting Machine) [5] classifier as the base estimator.

The best performance values observed on the test dataset for different metrics are: label-based micro average precision: 0.8265, label-based micro average recall: 0.8894, label-based micro average f1: 0.8568, label-based macro average precision: 0.7781, label-based macro average recall: 0.8022, label-based macro average f1: 0.7742, instance-based precision: 0.8589,

	Case Report	Diagnosis	Epidemic Forecasting	Mechanism	Prevention	Transmission	Treatment
Case Report	2063	0	0	0	0	0	0
Diagnosis	0	6193	8	746	733	287	3009
Epidemic Forecasting	0	8	645	4	422	60	7
Mechanism	0	746	4	4438	203	251	3433
Prevention	0	733	422	203	11102	612	695
Transmission	0	287	60	251	612	1088	141
Treatment	0	3009	7	3433	695	141	8717

Fig. 1. Label Co-occurrence Statistics.

instance-based recall: 0.9085, and instance-based f1: 0.883. The technique used for this prediction is described in Section 6.

The paper is organized as follows. Section 2 discusses the data preparation approaches and Section 3 describes the exploratory data analysis conducted on the training dataset which forms the basis of most of the conducted experiments. Section 4 describes the feature extraction approaches, and the details of the proposed models are given in Section 5. Section 6 provides an analysis of the obtained results. Section 7 concludes the paper.

II. Data Preparation

Different pre-processing steps have been applied to the given dataset to construct normalized text representations for different models, as discussed in later sections.

Each sample in the dataset is identified by its PMID, which is a unique identifier assigned to each PubMed record. Corresponding to each PMID, the title, abstract and keyword fields are concatenated to form a single text representation of the article. Based on the nature of the pre-processing done three different types of text files have been generated for each document. These are described as follows:

TXT1: In this representation, spacy pipeline with the abbreviation detector from scispacy [7] is applied to the text description. Thus, the abbreviations in the combined document texts are expanded.

TXT2: Here, scispacy's UMLS entity linker was used on TXT1 to perform text linking with UMLS knowledge base and replace the UMLS concepts identified from this raw text with their corresponding concept ids. An illustration of entity extraction from a title corresponding to PMID from the training dataset is given in Figures 2 – 5.

TXT3: Here the following pre-processing steps are applied to the TXT1 to obtain a new format of normalized text data.

a) Text is converted to lower case

b) Punctuations are removed

c) Standard English stop words are removed

A token consisting of two or more digits or alphabets is considered as a word while constructing a vocabulary. The size of the vocabulary is limited to top 20000 unigrams.

III. Exploratory Data Analysis

In this section we present the observations on label co-occurrence patterns from the training data. These observations consider the relative frequency with which different pairs of topics co-occur as a part of the label subset assigned to each article in the dataset. Figure 5 shows the label co-occurrence statistics for the training data. With each of the seven topics as pivots (bold-faced), clusters are observed with topics having similar values of co-occurrence frequency.

Fig. 2. Entities from the Title corresponding to PMID 32466289

	UMLS Concepts	Concept IDs
0	Cellular Metabolic Profiling	C1328813
1	Cells	C0007584
2	Infected	C0439663
3	Feline Infectious Peritonitis Virus	C0085305

Fig. 3. Identified UMLS concepts and concept Ids from the title

Fig. 4. Keywords corresponding to PMID 32466289

	UMLS Concepts	Concept IDs
0	cellular metabolism	C1524026
1	feline	C0325089
2	infectious peritonitis virus	C0085305
3	glutamine	C0017797
4	metabolic profiling	C1328813
5	phenotype	C0031437
6	microarray	C1709016

Fig. 5. Identified UMLS concepts and concept Ids from the keywords

Using this rule, the following clusters are used for partitioning the classifier: {**Case Report**}, {**Diagnosis**, Mechanism, Prevention}, {**Diagnosis**, Treatment}, {**Epidemic Forecasting**, Prevention}, {**Mechanism**, Treatment}, {**Mechanism**, Prevention, Transmission}, {**Prevention**, Diagnosis, Treatment}, {**Transmission**, Prevention}, {**Transmission**, Diagnosis, Mechanism, Treatment}, {**Treatment**, Diagnosis, Mechanism}

IV. Feature Extraction

Here we describe three different feature spaces, namely, FS1, FS2 and FS3, that we used for preparing the models.

FS1: Here, Term frequency-inverse document frequency (tf-idf) representation of the corpus is used as features. This representation has been extracted using genism [9] library, and applied to TXT1 and TXT2.

Given a corpus D with N documents and a vocabulary of size V, the term frequency of term t_i in a document d_j, is given by $tf(t_i, d_j) = \frac{f_{i,j}}{\sum_k f_{k,j}}$ where $f_{i,j}$ is the frequency of occurrence of the term t_i occurs in the document d_j.

The inverse document frequency is given by $\mathrm{idf}(t_i, D) = \log \frac{1}{n_i}$ where n_i is the number of documents in the corpus *D* containing the term t_i.

Tf-idf weight corresponding to the term t_i and document d_j is given by $\mathrm{tf}(t_i, d_j) \cdot \mathrm{idf}(t_i, D)$. We used a smoothened variation of idf, i.e., $\mathrm{idf}(t_i, D) = \log \frac{l+1}{n_i+1} + 1$ in order to take care of situations when $n_i = 0$ for some term.

FS2: Log entropy representation of the corpus is used as features. This representation is extracted using gensim library in Python. This technique is applied to TXT1 and TXT2.

Given a corpus *D* with N documents and a vocabulary of size V, the log entropy weight for term t_i in a document d_j is calculated as the product of a local weight (Lw) and a global weight (Gw), where:

$Lw_{ij} = \log(f_{i,j} + 1)$, where f_{ij} is the frequency of occurrence of the term t_i in the document d_j

$Gw_{ij} = 1 + \frac{\sum_j P_{i,j} \log(P_{i,j})}{\log(l+1)}$, where $P_{i,j} = \frac{f_{i,j}}{\sum_k f_{i,k}}$

Then the log entropy weights are given by $Lw_{ij} \cdot Gw_{ij}$

FS3: Here, contextual topic modelling (CTM) [10,11] is applied to pre-processed TXT3. CTM combines contextualized representations with neural topic models. It utilizes contextualized document embeddings from SBERT, to extend neural topic model ProdLDA. In particular, we used nli-roberta-base-v2 embedding from SBERT. Contextualized topic models use the pre-processed text for generating a BoW (Bag of Words) model, and use the raw text input for generating contextualized document embeddings.

In general, a topic modelling technique represents a document in the form of a probability distribution over latent topics. These are considered as features representing the training corpus.

V. Multi-Label Classifiers

Two types of classifiers are used for the experiments:

CLF1: Here a simple One-vs-the-rest classifier framework from scikit-learn library [12] is used with LGBM classifier as a base estimator, on feature space FS3.

LGBM is a Gradient Boosting Decision Tree algorithm that uses Gradient based One-Side Sampling to deal with large number of instances, and Exclusive Feature Bundling to deal with large number of features.

The One-vs-the-rest classifier is the same as Binary Relevance method for multilabel classification. This method fits a binary classifier corresponding to each label, learning presence or absence of that label in the label subset corresponding to each instance.

CLF2: Here, Network-based label space partition ensemble classifier from scikit-multilearn library [13] is used. This classifier divides the given multilabel problem into smaller sub-problems according to the specified label clusters.

To solve each of these multi-label classification sub-problems, Random k-Labelsets (RAKEL) classifier is used. RAKEL is an ensemble algorithm which constructs each component of the ensemble by considering a small random subset of labels (size k), and learn a single label classifier using base estimator for prediction of each instance in the power set of this subset of labels. In our experiments we used LGBM classifier as the base estimator for RAKEL. This combined classifier (viz. CLF2) has been applied to the given multi-label classification problem on the clusters defined in Section 3.

VI. Results And Discussion

The metrics used for evaluation are:

LP_Micro, LR_Micro, LF1_Micro: Label-based micro average precision, recall and F1 score, respectively.

LP_Macro, LR_Macro, LF1_Macro: Label-based macro average precision, recall and F1 score, respectively.

IP, IR, IF1: Instance-based precision, recall, and F1 score, respectively.

The five sets of experiments which performed best on the development dataset were used to make predictions on the test set.

The experiments are named using the following technique, '<Data Preparation Technique> + <Feature Extraction Technique> + <Classifier>'. The results obtained on the test set are given in Table 1.

The best performing method in majority of the metrics is the method using tf-idf representation of raw text as features, with RAKEL classifier and LGBM base estimator.

We observe that in case of most of the metrics, tf-idf representation outperforms log-entropy representation.

Baseline model used for evaluation of this task is ML-NET [4]. The performance metrics for the baseline model are given in Table 2.

VII. Conclusion

The aim of the present work is to develop a model for multi-label classification. Experiments were conducted using two types of approaches. The first one is a partition based approach, which divides the given multi-label problem into sub-problems that are solved using the Random k-labelsets algorithm. This approach uses tf-idf representation and log-entropy representation as features. In the second approach uses document-topic distribution from Contextual topic

TABLE I. RESULTS OF DIFFERENT EXPERIMENTS

(THE ROW HEADERS ARE EXPLAINED IN SECTIONS II, IV AND V)

Algorithm	Metrics								
	LP_Micro	***LR_Micro***	***LF1_Micro***	***LP_Macro***	***LR_Macro***	***LF1_Macro***	***IP***	***IR***	***IF1***
TXT1+FS1+CLF2	0.8265	**0.8894**	**0.8568**	**0.7781**	**0.8022**	**0.7742**	**0.8589**	**0.9085**	**0.8830**
TXT1+FS2+CLF2	0.8166	0.8844	0.8492	0.7652	0.7986	0.7624	0.8508	0.9028	0.8760
TXT2+FS1+CLF2	0.8206	0.8473	0.8337	0.784	0.7315	0.7445	0.8372	0.8622	0.8495
TXT2+FS2+CLF2	0.8089	0.8465	0.8273	0.7543	0.7332	0.7217	0.8345	0.8660	0.8500
TXT3+FS3+CLF1	**0.8419**	0.7572	0.7973	0.7645	0.6323	0.6717	0.8112	0.784	0.7974

modelling as features, which is then solved using a binary relevance classifier with LGBM classifier as base estimator.

Data for the classification has been prepared by using different pre-processing schemes applied to the original text representation. One of these approaches use UMLS concepts extracted from the text, but it is observed that the models based on raw text representation, i.e. the ones without UMLS concepts, do exhibit superior performance.

The partition-based method on raw text representation with tf-idf features performed the best amongst the conducted experiments. Moreover, they showed significant improvement upon the baseline performance in majority of the evaluation metrics.

VIII. REFERENCES

1. Chen, Q., Allot, A., Leaman, R., Doğan, R.I., and Lu, Z., (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.

2. Chen, Q., Allot, A. and Lu, Z., (2020). Keep up with the latest coronavirus research. Nature, 579(7798), pp.193-194.

3. Chen, Q., Allot, A. and Lu, Z., (2021). LitCovid: an open database of COVID-19 literature. Nucleic acids research, 49(D1), pp.D1534-D1540.

4. Du, J., Chen, Q., Peng, Y., Xiang, Y., Tao, C. and Lu, Z., (2019). ML-Net: multi-label classification of biomedical texts with deep neural networks. Journal of the American Medical Informatics Association, 26(11),pp.1279-1285.

5. Ke, G., Meng, Q., Finley, T., et al., (2017). LightGBM: a highly efficient gradient boosting decision tree. Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, Curran Associates Inc., Long Beach, California, USA, pp. 3149–3157.

6. Tsoumakas, G., Katakis, I. and Vlahavas, I., (2011). Random k-Labelsets for Multilabel Classification. IEEE Transactions on Knowledge and Data Engineering, 23, 1079–1089.

7. Neumann, M., King, D., Beltagy, I., et al., (2019). ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. Proceedings of the 18th BioNLP Workshop and Shared Task, Association for Computational Linguistics, Florence, Italy, pp. 319–327.

8. Bodenreider, O., (2004). The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res, 32, D267-270.

9. Řehůřek, R. and Sojka, P., (2010). Software Framework for Topic Modelling with Large Corpora. Software Framework for Topic Modelling with Large Corpora; University of Malta, (2010) .

10. Bianchi, F., Terragni, S. and Hovy, D., (2021). Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Association for Computational Linguistics, Online, pp. 759–766.

11. Bianchi, F., Terragni, S., Hovy, D., et al., (2021). Cross-lingual Contextualized Topic Models with Zero-shot Learning. Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Association for Computational Linguistics, Online, pp. 1676–1683.

12. Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011) Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

13. Szymański, P. and Kajdanowicz, T. (2018) A scikit-based Python environment for performing multi-label classification. arXiv:1702.01460 [cs].

TABLE II. BASELINE PERFORMANCE

Algorithm	Metrics								
	LP_Micro	***LR_Micro***	***LF1_Micro***	***LP_Macro***	***LR_Macro***	***LF1_Macro***	***IP***	***IR***	***IF1***
ML-NET	0.8756	0.8142	0.8437	0.8364	0.7309	0.7655	0.8849	0.8514	0.8678