












# Multi-label classification for biomedical literature: an overview of the BioCreative VII LitCovid Track for COVID-19 literature topic annotations

Qingyu Chen <sup>1,†</sup>, Alexis Allot <sup>1,†</sup>, Robert Leaman<sup>1</sup>, Rezarta Islamaj<sup>1</sup>, Jingcheng Du <sup>2</sup>, Li Fang <sup>3</sup>, Kai Wang<sup>3,4</sup>, Shuo Xu <sup>5</sup>, Yuefu Zhang<sup>5</sup>, Parsa Bagherzadeh<sup>6</sup>, Sabine Bergler<sup>6</sup>, Aakash Bhatnagar<sup>7</sup>, Nidhir Bhavsar<sup>7</sup>, Yung-Chun Chang <sup>8</sup>, Sheng-Jie Lin<sup>8</sup>, Wentai Tang<sup>9</sup>, Hongtong Zhang<sup>9</sup>, Ilija Tavchioski<sup>10,11</sup>, Senja Pollak<sup>11</sup>, Shubo Tian <sup>12</sup>, Jinfeng Zhang <sup>12</sup>, Yulia Otmakhova<sup>13</sup>, Antonio Jimeno Yepes<sup>14</sup>, Hang Dong<sup>15</sup>, Honghan Wu <sup>16</sup>, Richard Dufour<sup>17</sup>, Yanis Labrak<sup>18</sup>, Niladri Chatterjee<sup>19</sup>, Kushagri Tandon<sup>19</sup>, Fréjus A. A. Laleye<sup>20</sup>, Loïc Rakotoson <sup>20</sup>, Emmanuele Chersoni<sup>21</sup>, Jinghang Gu<sup>21</sup>, Annemarie Friedrich<sup>23</sup>, Subhash Chandra Pujari<sup>22,23</sup>, Mariia Chizhikova<sup>24</sup>, Naveen Sivadasan<sup>25</sup>, Saipradeep VG<sup>25</sup> and Zhiyong Lu <sup>1,\*</sup>

<sup>1</sup>National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, MD, Bethesda 20892, USA

<sup>2</sup>School of Biomedical Informatics, UT Health, TX, Houston 77030, USA

<sup>3</sup>Raymond G. Perelman Center for Cellular and Molecular Therapeutics, Children's Hospital of Philadelphia, Philadelphia, PA, USA

<sup>4</sup>Department of Pathology and Laboratory Medicine, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA

<sup>5</sup>College of Economics and Management, Beijing University of Technology, Beijing, QC, China

<sup>6</sup>CLaC Labs, Concordia University, Montreal, Canada

<sup>7</sup>Navrachana University, Vadodara, India

<sup>8</sup>Graduate Institute of Data Science, Taipei Medical University, Taipei, Taiwan

<sup>9</sup>College of Computer Science and Technology, Dalian University of Technology, Dalian, China

<sup>10</sup>Computer and Information Science, University of Ljubljana, Ljubljana, Slovenia

<sup>11</sup>Jožef Stefan Institute, Ljubljana, Slovenia

<sup>12</sup>Department of Statistics, Florida State University, Tallahassee, FL, USA

<sup>13</sup>School of Computing and Information Systems, University of Melbourne, Melbourne, AU-VIC, Australia

<sup>14</sup>School of Computing Technologies, RMIT University, Melbourne, AU-VIC, Australia

<sup>15</sup>Centre for Medical Informatics, Usher Institute, University of Edinburgh, Edinburgh, UK

<sup>16</sup>Institute of Health Informatics, University College London, London, UK

<sup>17</sup>LS2N, Nantes University, Nantes, France

<sup>18</sup>LIA, Avignon University, Avignon, France

<sup>19</sup>Department of Mathematics, Indian Institute of Technology Delhi, New Delhi, India

<sup>20</sup>Opscidia, Paris, France

<sup>21</sup>Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong, China

<sup>22</sup>Institute of Computer Science, Heidelberg University, Heidelberg, Germany

<sup>23</sup>Bosch Center for Artificial Intelligence, Renningen, Germany

<sup>24</sup>SINAI Group, Department of Computer Science, Advanced Studies Center in ICT (CEATIC), Universidad de Jaén, Jaén, Spain

<sup>25</sup>TCS Research, Life Sciences, Hyderabad, India

\*Corresponding author: Email: [zhiyong.lu@nih.gov](mailto:zhiyong.lu@nih.gov)

<sup>†</sup>These authors contributed equally to this work.

Citation details: Chen, Q., Allot, A., Leaman, R. *et al.* Multi-label classification for biomedical literature: an overview of the BioCreative VII LitCovid Track for COVID-19 literature topic annotations. *Database* (2022) Vol. 2022: article ID baac069; DOI: <https://doi.org/10.1093/database/baac069>

## Abstract

The coronavirus disease 2019 (COVID-19) pandemic has been severely impacting global society since December 2019. The related findings such as vaccine and drug development have been reported in biomedical literature—at a rate of about 10 000 articles on COVID-19 per month. Such rapid growth significantly challenges manual curation and interpretation. For instance, LitCovid is a literature database of COVID-19-related articles in PubMed, which has accumulated more than 200 000 articles with millions of accesses each month by users worldwide. One primary curation task is to assign up to eight topics (e.g. Diagnosis and Treatment) to the articles in LitCovid. The annotated topics have been widely used for navigating the COVID literature, rapidly locating articles of interest and other downstream studies. However, annotating the topics

has been the bottleneck of manual curation. Despite the continuing advances in biomedical text-mining methods, few have been dedicated to topic annotations in COVID-19 literature. To close the gap, we organized the BioCreative LitCovid track to call for a community effort to tackle automated topic annotation for COVID-19 literature. The BioCreative LitCovid dataset—consisting of over 30 000 articles with manually reviewed topics—was created for training and testing. It is one of the largest multi-label classification datasets in biomedical scientific literature. Nineteen teams worldwide participated and made 80 submissions in total. Most teams used hybrid systems based on transformers. The highest performing submissions achieved 0.8875, 0.9181 and 0.9394 for macro-F1-score, micro-F1-score and instance-based F1-score, respectively. Notably, these scores are substantially higher (e.g. 12%, higher for macro F1-score) than the corresponding scores of the state-of-art multi-label classification method. The level of participation and results demonstrate a successful track and help close the gap between dataset curation and method development. The dataset is publicly available via <https://ftp.ncbi.nlm.nih.gov/pub/lu/LitCovid/biocreative/> for benchmarking and further development.

**Database URL:** <https://ftp.ncbi.nlm.nih.gov/pub/lu/LitCovid/biocreative/>

## Introduction

The rapid growth of biomedical literature poses a significant challenge for manual curation and interpretation [1–3]. This challenge has become more evident during the coronavirus disease 2019 (COVID-19) pandemic: the number of COVID-19-related articles in the literature is growing by about 10 000 articles per month; the median number of new articles per day since May 2020 is 319, with a peak of over 2500, and this volume accounts for over 7% of all PubMed articles [4].

In response, LitCovid [5, 6], the first-of-its-kind COVID-19-specific literature resource, has been developed for tracking and curating COVID-19-related literature. Every day, it triages COVID-19-related articles from PubMed, categorizes the articles into research topics (e.g. prevention measures) and recognizes and standardizes the entities (e.g. vaccines and drugs) mentioned in each article. The collected articles and curated data in LitCovid are freely available. Since its release, LitCovid has been widely used with millions of accesses each month by users worldwide for various information needs, such as evidence attribution, drug discovery and machine learning [6].

Initially, data curation in LitCovid was done manually with little machine assistance. The rapid growth of the COVID-19 literature significantly increased the burden of manual curation, especially for topic annotations [6]. Topic annotation in LitCovid is a standard multi-label classification task that assigns one or more labels to each article. A set of eight topics are selected for annotation based on topic modeling and discussions with physicians aiming to understand COVID-19, such as the Transmission topic, which describes the characteristics and modes of COVID-19 transmissions. The annotated topics have been demonstrated to be effective for information retrieval and have been widely used in many downstream applications. Topic-related searching and browsing accounts for ~20% of LitCovid user behaviors, making it the second-most-used feature in LitCovid [6]. The topics have also been used in downstream studies such as citation analysis and knowledge network generation [7–9]. Figure 1 shows the characteristics of topic annotations in LitCovid.

However, annotating topics in LitCovid has been a primary bottleneck for manual curation. Compared to other curation tasks in LitCovid (document triage and entity recognition), topic annotation is more difficult due to the requirement of interpretation of the biomedical literature and assignment of up to eight topics. As an example of the language variation that must be addressed, we provide the following five sentence snippets reflecting the treatment topic: (i) ‘...as a management option for COVID-19-associated diarrhea...’ (PMID34741071), (ii) ‘...modulating these factors may impact in guiding the success of vaccines and clinical outcomes in COVID-19 infections...’

(PMID34738147), (iii) ‘...lung ultrasound abnormalities are prevalent in patients with severe disease, RV involvement seems to be predictive of outcomes...’ (PMID34737535), (iv) ‘...common and virus-specific host responses and vRNA-associated proteins that variously promote or restrict viral infection...’ (PMID34737357) and (v) ‘...the unique ATP-binding pockets on NTD/CTD may offer promising targets for design of specific anti-SARS-CoV-2 molecules to fight the pandemic...’ (PMID34734665). Although these sentence snippets all describe treatment-related information, they use rather different vocabularies and structures. While automatic approaches have been developed to assist manual curation in LitCovid, the evaluations show that the automatic topic annotation tool has an F1-score of 10% lower than the tools assisting other curation tasks in LitCovid [6]. Increasing the accuracy of automated topic prediction in COVID-19-related literature would be a timely improvement beneficial to curators, biomedical researchers and healthcare professionals.

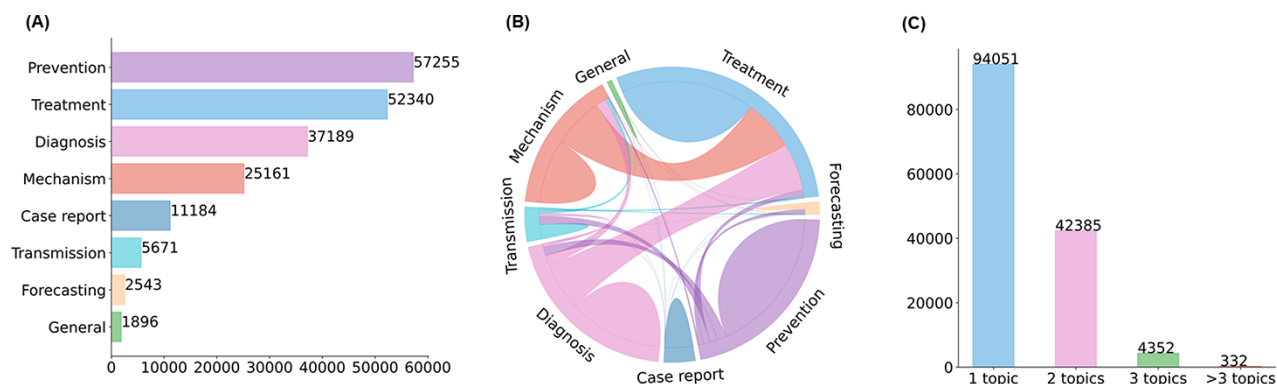
To this end, we organized the BioCreative LitCovid track to call for a community effort to tackle automated topic annotation for COVID-19 literature. BioCreative, established in 2003, is the first and longest-running community-wide effort for assessing biomedical text-mining methods [10]. Previous BioCreative challenges have successfully organized tracks on a range of biomedical text-mining applications such as relation extraction [11] and entity normalization [12].

This article provides an extended overview from [13] on the BioCreative LitCovid track. It substantially describes (i) the dataset annotation characteristics, (ii) detailed methods from the participating teams and (iii) in-depth evaluation results. Overall, 19 teams submitted 80 runs, and ~75% of the submissions had better performance than the baseline method [14]. The dataset and evaluation scripts are available via <https://ftp.ncbi.nlm.nih.gov/pub/lu/LitCovid/biocreative/> and [https://github.com/ncbi/biocreative\\_litcovid](https://github.com/ncbi/biocreative_litcovid), respectively. We encourage further work to develop multi-label classification methods for biomedical literature.

## Dataset, baselines and evaluation measures

### The overall LitCovid curation pipeline

The LitCovid curation pipeline has three primary modules: (i) document triage, identifying COVID-19-related articles from new articles in PubMed, (ii) topic classification, assigning up to eight topics to the COVID-19-related articles (i.e. a multi-label classification task) and (iii) entity recognition, extracting chemicals and locations mentioned in these articles. Initially, the curation was done manually with little machine assistance by two (part-time) human curators with a background in biomedical data sciences. As the outbreak evolved, we developed automated approaches to support manual curation



**Figure 1.** Characteristics of topic annotations in LitCovid up to Feb 2022. (A) shows the frequencies of topics; (B) demonstrates topic co-occurrences and (C) illustrates the distributions of the number of topics assigned per document.

and maximize curation productivity to keep up with the rapid literature growth. The detailed implementation and evaluation of the automated approaches are fully described in the description of the LitCovid resource [6]. In summary, all automated methods were evaluated before first use and have been improved continuously. The evaluations demonstrated that automated methods can achieve exceptionally high performance for document triage and entity recognition (e.g. the F1-scores were 0.99 and 0.94 for document triage and entity recognition, respectively). In contrast, the F1-score of the topic classification was 0.80, largely due to the complexity of the multi-label classification task, which assigns up to eight topics. We therefore organized this to call for a community effort to tackle automated topic annotation for COVID-19 literature.

### Topic annotations in LitCovid

The topic annotation step assigns up to eight topics to the COVID-19-related articles:

1. **Case Report:** descriptions of specific patient cases related to COVID-19,
2. **Diagnosis:** COVID-19 assessment through symptoms, test results and radiological features for COVID-19,
3. **Epidemic Forecasting:** estimation on the trend of COVID-19 spread and related modeling approach,
4. **General Information:** COVID-19-related brief reports and news,
5. **Mechanism:** underlying cause(s) of COVID-19 infections and transmission and possible drug mechanism of action,
6. **Prevention:** prevention, control, mitigation and management strategies,
7. **Transmission:** characteristics and modes of COVID-19 transmissions,
8. **Treatment:** treatment strategies, therapeutic procedures and vaccine development for COVID-19.

Note that by design Case Report and General Information are singleton topics, i.e. not co-assigned with other topics. This is due to their broad scope, e.g. a case report typically also contains diagnostic information.

Topics are annotated mainly based on titles and abstracts of the papers; the curators may also look for other information

such as full-text and Medical Subject Headings (MeSH) when needed. Previous studies have shown that many COVID-19 articles published in PubMed without abstract information are not descriptions of formal research studies but rather commentary or perspective [15]. We also find that automatic topic annotation methods achieve 10% higher F1-score on articles with abstracts available [6]. Since late August 2020, we have prioritized annotating topics for the articles with abstracts available in PubMed, when the number of daily new articles reached a record high of over 2500.

### Dataset characteristics

Table 1 summarizes the dataset characteristics in terms of the scale of the dataset, labels and annotators. It also compares the dataset with representative counterparts. There are only a few existing multi-label classification datasets for biomedical scientific literature, and their size is relatively small. The Hallmarks of Cancer dataset [16] has been widely used for multi-label classification methods, which has about ~1600 documents. Another dataset on chemical exposure assessment [17] has ~3700 documents. In contrast, The BioCreative LitCovid dataset has ~34 000 documents in total, which is nearly 10 times larger. The training, development and testing sets contain 24 960, 6239 and 2500 articles in LitCovid, respectively. Table 2 shows the detailed topic distributions of the dataset. The topics were assigned using the above annotation approach consistently. All the articles contain both titles and abstracts available in PubMed and have been manually reviewed by curators. The only difference is that the datasets do not contain the General Information topic since the priority is given to the articles with abstracts available in PubMed. The training and development datasets were made available on 15 June 2021, to all participant teams. The testing set contains held-out articles added to LitCovid from June 16 to 22 August 2021. Using incoming articles to generate the testing set facilitates the evaluation of the generalization capability of automatic tools.

In addition, most existing multi-label datasets on biomedical literature were annotated by a single curator, which does not allow inter-annotator agreement to be measured. A random sample of 200 articles in LitCovid was used to measure inter-annotator agreement, and two curators annotated each article independently. Table 3 shows that the micro-average of Pearson correlation of the curators across the seven topics is 0.78, which can be interpreted as ‘strong correlation’ [18].

**Table 1.** BioCreative LitCovid dataset characteristics in comparison with representative multi-label classification datasets on biomedical scientific literature

	Dataset scale				Label scale			Annotator scale
	Total documents	Train	Valid	Test	Total labels	Avg. labels per doc	Unique labels	Annotators
Hallmarks of Cancer [16]	1580	1108	157	315	2469	1.56	10	1
Chemical Exposure [17]	3661	–	–	–	21 233	5.80	32	1
BioCreative LitCovid (ours)	33 699	24 960	6239	2500	46 368	1.38	7	2

Note that the Chemical Exposure dataset does not provide dataset splits.

**Table 2.** Detailed topic annotation characteristics

	Train		Valid		Test		All	
	#Articles	Label (%)	#Articles	Label (%)	#Articles	Label (%)	#Articles	Label (%)
Case Report	2063	(8.27%)	482	(7.73%)	197	(7.88%)	2742	(8.14%)
Diagnosis	6193	(24.81%)	1546	(24.78%)	722	(28.88%)	8461	(25.11%)
Epidemic Forecasting	645	(2.58%)	192	(3.08%)	41	(1.64%)	878	(2.61%)
Mechanism	4438	(17.78%)	1073	(17.2%)	567	(22.68%)	6078	(18.04%)
Prevention	11 102	(44.48%)	2750	(44.08%)	926	(37.04%)	14 778	(43.85%)
Transmission	1088	(4.36%)	256	(4.1%)	128	(5.12%)	1472	(4.37%)
Treatment	8717	(34.92%)	2207	(35.37%)	1035	(41.4%)	11 959	(35.49%)

Note that the General Information topic is excluded as the annotation priority is given to the articles with abstracts available in PubMed.

**Table 3.** Inter-annotator agreement on a random sample of 200 articles

Topic	Size (percentage)	Pearson correlation
Case Report	15 (7.50%)	0.90
Diagnosis	37 (18.50%)	0.71
Epidemic Forecasting	5 (2.50%)	0.51
Mechanism	35 (17.50%)	0.72
Prevention	94 (47.00%)	0.84
Transmission	4 (2.00%)	0.66
Treatment	66 (33.00%)	0.77
Macro-average	–	0.73
Micro-average	–	0.78

Note that the General Information topic is excluded as the annotation priority is given to the articles with abstracts available in PubMed.

The distribution of the topics in the random sample is also consistent with that of the entire dataset. Given the scale of the dataset, the curator each annotated half of the remaining dataset and discussed difficult cases together.

## Baseline method

We chose Machine Learning (ML)-Net [14] as the baseline method. ML-Net is a deep learning framework specifically for multi-label classification tasks for biomedical literature. It has achieved favorable state-of-the-art performance in a few biomedical multi-label text classification tasks, and its source code is publicly available [14]. ML-Net first maps texts into high-dimensional vectors through deep contextualized word representations (ELMo) [19] and then combines a label prediction network and label count prediction to infer an optimal set of labels for each document. We ran ML-Net with ten different random seeds and reported the median performance.

## Evaluation measures

Evaluation measures for multi-label classification tasks can be broadly divided into two groups: (i) label-based measures, which evaluate the classifier's performance on each label,

and (ii) instance-based measures (also called example-based measures), which aim to evaluate the multi-label classifier's performance on each test instance [20–22]. Both groups have unique strengths and complement each other: label-based measures quantify the effectiveness of each individual label, whereas instance-based measures quantify the effectiveness of instances which may contain multiple labels. We employed representative metrics from both groups to provide a broader evaluation of the performance. Specifically, for label-based measures, we calculated macro- and micro-averages on precision, recall and F1-score. The macro-average computes the arithmetic average by considering all the topics equally regardless of the number of instances per class, whereas the micro-average computes the weighted average according to the number of instances. For instance-based measures, we calculated instance-based precision, recall and F1-score. Out of these nine metrics, we focus on the three F1-scores because these aggregate both precision and recall.

## Results and discussion

### Participating teams

Table 4 provides details on the participating teams and their number of submissions. Each team is allowed to submit up to five test set predictions. Overall, 19 teams submitted 80 valid testing set predictions in total.

### System descriptions

Out of 19 teams, 17 teams agreed to participate in the track overview and described their approaches. Table 5 summarizes their methods and associated performance. The full detail is further provided in Table S1. Overall, we notice that the transformer approach has been used extensively: 14 out of the 17 teams (82.3%) used transformers purely (nine teams) and a combination of transformers and other traditional deep learning approaches (five teams). In contrast, only

**Table 4.** Team participation details, ordered alphabetically by team name

Team name	Team affiliation	Submissions
Bioformer	Children's Hospital of Philadelphia	5
BJUT-BJFU	Beijing University of Technology and Beijing Forestry University	5
CLaC	Concordia University	4
CUNI-NU	Navrachana University and Charles University	5
DonutNLP	Taipei Medical University, Taipei Medical University Hospital and National Tsing Hua University	5
DUT914	Dalian University of Technology	3
E8@IJS	Jozef Stefan Institute	3
ElsevierHealth-Sciences	Elsevier	1
FSU2021	Florida State University	5
itcc	University of Melbourne and RMIT University	4
KnowLab	University of Edinburgh and University College London	5
LIA/LS2N	Avignon Université	4
LRL_NC	Indian Institute of Technology Delhi	5
Opscidia	Opscidia	5
PIDNA	Roche Holding Ltd	3
polyu_cbsnlp	The Hong Kong Polytechnic University and Tencent AI Lab	5
robert-nlp	Bosch Center for Artificial Intelligence and Bosch Global	5
SINAI	Universidad de Jaén	4
TCSR	Tata Consultancy Services	4

two teams used deep learning approaches besides transformers only and two teams used machine learning approaches only. This is different from previous BioCreative challenge tasks, where most teams used machine learning approaches or a combination of machine learning and deep learning techniques [11, 23–26]. In addition, of the 14 teams using the transformer approach, seven teams (50%) proposed innovative methods beyond the default approach (fine-tuning the transformers). For instance, the Bioformer team proposed a lightweight transformer architecture, which reduces the number of parameters by two-thirds (the detail is summarized in [27]); the DUT914 team proposed an enhanced transformer model, which learns the correlations between labels for the multi-label classification task (the detail is summarized in [28]). Such innovative approaches demonstrated superior performance and achieved top-ranked results. In addition, six teams (35%) used additional data (beyond titles and abstracts) for training the models, including metadata (e.g. paper types and journals), entity annotations (e.g. Unified Medical Language System (UMLS) [29] and DrugBank [30]) and synonyms (e.g. WordNet [31]).

#### Bioformer team [27]

We performed topic classification using three BERT models: BioBERT [32], PubMedBERT [33] and Bioformer (<https://github.com/WGLab/bioformer/>). For BioBERT, we used BioBERT<sub>Base-v1.1</sub>, which is the version described in the publication [32]. PubMedBERT has two versions: one version was pretrained on PubMed abstracts (denoted by PubMedBERT<sub>Ab</sub> in this study) and the other version was pretrained on PubMed abstracts plus PMC full texts (denoted

by PubMedBERT<sub>AbFull</sub>). We used Bioformer<sub>gL</sub>, which is pretrained on PubMed abstracts and 1 million PMC full-text articles for 2 million steps. We formulated the topic classification task as a sentence pair classification problem where the title is the first sentence and the abstract is the second sentence. The input is represented as '[CLS] title [SEP] abstract [SEP]'. The representation of the [CLS] token in the last layer was used to classify the relations. We utilized the sentence classifier in the transformers python library to fine-tune the models. We treated each topic independently and fine-tuned seven different models (one per topic). We fine-tuned each BERT model on the training dataset for three epochs. The maximum input sequence length was fixed to 512. We selected a batch size of 16 and a learning rate of  $3e-5$ .

#### BJUT-BJFU team [34]

We combined the training and development sets to create our training set, which we further grouped into 10 disjoint subsets with nearly equal size and similar label distribution using the stratification method in Sechidis *et al.* [35]. Our method takes advantage of four powerful deep learning models: FastText [36], Text Recurrent Convolutional Neural Network (TextR-CNN) [37], Text Convolutional Neural Network (TextCNN) [38] and Transformer [39]. We also consider the correlations among labels [40].

#### CLaC team [41]

We used a multi-label classification approach, where a base network (shared by several classifiers) is responsible for representation learning for all classes. Although the classes might be related, different classes often require focus on different parts of the input. To allow a differential focus on the input, we used the multi-input Recurrent Independent Mechanisms (RIM) model [42] with seven class modules, one for each class, each using ClinicalBERT [43] as input. We also used a gazetteer module for leveraging annotations from DrugBank [30] and MeSH [44]. The modules sparsely interact with one another through an attention bottleneck, enabling the system to achieve compositional behavior. The proposed model improves all classes, especially the two least frequent classes, Transmission and Epidemic Forecasting. Moreover, the functionality of the modules is transparent for inspection [45].

#### CUNI-NU team [46]

Our approach implemented the Specter model [47], which incorporates SciBERT [48] to produce the document-level embedding using citation-based transformers. SciBERT can decipher the dense biomedical vocabulary in the COVID-19 literature, making it a valuable choice. Furthermore, we used a dual-attention module [39], consisting of two self-attention layers applied to the embeddings in sequential order. These self-attention layers allow each input to establish relationships with other instances. To obtain unique vectors, i.e. query (Q), key (K) and value (V), three individually learned matrices are multiplied with the input vector. A single self-attention layer can learn the relationship between contextual semantics and sentimental tendency information. The dual self-attention mechanism helps retain more information from the sentence and thus generates a more representative feature vector. However, the dual self-attention mechanism can only generate relationships among the input instances while

**Table 5.** Systems and performance. The systems are categorized in terms of additional training data and knowledge sources, backbone models and methods. The best performance in terms of each metric is also reported

Team name	Systems		Best performance		
	Additional training data and knowledge sources	Models and methods	Micro-F1	Macro-F1	Instance F1
Bioformer	–	BioBERT, PubMedBERT and Bioformer	0.9181	0.8875	0.9334
BJUT-BJFU	–	FastText, TextRCNN, TextCNN, Transformer and correlation learning	0.8556	0.7847	0.8701
CLaC	DrugBank and MeSH	Multi-input RIM model and ClinicalBERT	0.8897	0.8487	0.9102
CUNI-NU	–	SciBERT, dual-attention modules and LWAN	0.8959	0.8673	0.9153
DonutNLP	–	BioBERT and ensemble learning	0.9174	0.8754	0.9346
DUT914	–	BioBERT and label feature enhancement module	0.9175	0.8760	0.9394
E8@IJS	–	AutoBOT and doc2vec	0.8430	0.7382	0.8518
FSU2021	–	PubMedBERT and multi-instance learning	0.9067	0.8670	0.9247
ITTC	–	SVM, SciBERT, Specter, BioELECTRA and ensemble learning	0.9000	0.8669	0.9185
KnowLab	Back translation (to German), keywords, journals, UMLS, MeSH, SJR journal categories	BlueBERT-Base, PubMedBERT, JMAN, HLAN, HA-GRU, HAN, CNN, LSTM and ensemble learning	0.8932	0.8601	0.9169
LIA/LS2N	–	TARS transformer, few-shot learning and TF-IDF	0.8830	0.8366	0.9094
LRL_NC	–	Co-occurrence learning, TF-IDF and LGBM	0.8568	0.7742	0.8830
Opscidia	–	BERT, data augmentation and ensemble learning	0.9135	0.8824	0.9296
polyu_cbsnlp	MeSH	BioBERT-Base, BioBERT-Large, PubMedBERT, CovidBERT, BioELECTRA, BioM-ELECTRA, BioMed_RoBERTa and ensemble learning	0.9139	0.8749	0.9319
robert-nlp	Publication type, keywords and journals	SciBERT	0.9032	0.8655	0.9251
SINAI	Synonyms from WordNet	Logistic regression and TF-IDF	0.8254	0.7643	0.8086
TCSR	Biomedical entities	BioBERT and ensemble learning	0.8495	0.7896	0.8845

completely discarding the output. A Label-Wise-Attention-Network (LWAN) [49] is used to improve the results further and overcome the limitation of dual attention. LWAN provides attention to each label in the dataset and improves individual word predictability by paying special attention to the output labels. It uses attention to allow the model to focus on specific words in the input rather than memorizing the essential features in a fixed-length vector. Label-wise attention mechanism repeatedly applies attention  $L$  (number of labels) times, where each attention module is reserved for a specific label. Weighted binary cross-entropy is used as a loss function. This loss function was most appropriate as it gives equal importance to the different classes during training, which was necessary due to the significant imbalance in the data. Thus, this approach overcame the significant imbalance among class labels and attained extensive results on labels like Case study, Epidemic Forecasting, Transmission and Diagnosis.

#### DonutNLP team [50]

We proposed a BERT-based Ensemble Learning Approach to predict topics for the COVID-19 literature. To select the best BERT model for this task, we conducted experiments estimating the performance of several BERT models using training data. The results demonstrate that BioBERTv1.2 achieved the best performance out of all models. We then used ensemble learning with a majority voting mechanism to integrate multiple BioBERT models, which are selected by the results of  $k$ -fold cross-validation. Finally, our proposed method can achieve remarkable performances on the official dataset with

precision, recall and F1-score of 0.9440, 0.9254 and 0.9346, respectively.

#### DUT914 team [28]

We designed a feature enhancement approach to address the problem of insufficient features in medical datasets. First, we extract the article titles and the abstracts from the dataset. Then the article title and the abstract are concatenated as the first input part. We only take the article titles as the second input part. Additionally, we count the distribution of labels in the training set and design a tag association matrix based on the distribution. Second, we process the features to achieve feature equalization. The first input part is tokenized and then encoded by the pretrained model BioBERT [32]. The second input part is embedded randomly. Then, we concatenate the processed features and the tokenization of the title to obtain the equalized features. Finally, we design a feature enhancement module to integrate the previously obtained label features into the model. We multiply the equalized features by the label matrix to obtain the final output vector used for classification.

#### E8@IJS team [51]

Our approach [51] used the automated Bag-Of-Tokens (autoBOT) system by Škrlj *et al.* [52] with some task-specific modifications. The main idea of the autoBOT system is representation evolution by learning the weights of different representations, including token, sub-word and sentence-level

(contextual and non-contextual) features. The system produces a final representation that is suitable for the specific task.

First, we transformed the multi-label classification task into a binary classification problem by treating each assignable topic as a binary classification. Next, we developed three configurations of the autoBOT system. The first configuration Neural includes two doc2vec-based latent representations, each with a dimensionality of 512. The second configuration, Neurosymbolic-0.1, includes both symbolic and sub-symbolic features, where the symbolic features include features based on words, characters, part-of-speech tags and keywords; the dimension of the symbolic feature subspaces is 5120. The third configuration, Neurosymbolic-0.02, has symbolic and sub-symbolic features, the same as the second configuration, but the dimensionality of the symbolic feature subspaces is 25 600.

Even if the organizers' baseline model [14] has better performance in most of the metrics, the Neurosymbolic-0.1 configuration of the autoBOT system achieves label-based micro- and macro-precision of 0.8930 and 0.9175, respectively, by which it outperforms the baseline system (for 8% points in terms of macro-precision). Moreover, by our results of label-based F1-score (micro) of 0.8430 (Neurosymbolic-0.02 configuration) and F1-score (macro) of 0.7382 (Neural configuration), the system has results comparable to the state-of-the-art baseline system (cca. 2% below), which indicates that autoML is a promising path for future work.

#### FSU2021 team [53]

In our participation in the BioCreative VII LitCovid track, we evaluated several deep learning models built on PubMedBERT, a pretrained language model, with different strategies to address the challenges of the task. Specifically, we used multi-instance learning to deal with the large variation in the lengths of the articles and used the focal loss function to address the imbalance in the distribution of different topics. We also used an ensemble strategy to achieve the best performance among all the models. Test results of our submissions showed that our approach achieved a satisfactory performance with an F1-score of 0.9247, which is significantly better than the baseline model (F1-score: 0.8678) and the average of all the submissions (F1-score: 0.8931).

#### ITTC team [54]

Team ITTC combined traditional bag-of-words classifiers such as their implementation of MTI ML (a linear Support Vector Machine (SVM) model using gradient descent and the modified Huber loss [55, 56], available at <https://github.com/READ-BioMed/MTIMLExtension>) and neural models including SciBERT [48], Specter [47] and BioELECTRA [57]. We combined these into two ensemble methods: averaging across the results of SciBERT, MTI ML and Specter, on the one hand, and taking the maximum of scores assigned by SciBERT and MTI ML, on the other. The reason for such ensembles was that SciBERT tended to give high scores to well-represented categories such as Treatment while assigning scores close to zero for weaker classes such as Transmission, so its performance varied greatly depending on the composition of the test set. Conversely, Specter and

MTI ML were more conservative but assigned more scores close to 0.5 even for underrepresented labels, which improved precision for difficult categories. The ensemble based on the maximum value proved to be an effective strategy for recall, while averaging improved precision, especially for underrepresented and challenging categories, which led to very strong macro-precision results.

#### KnowLab team [58]

KnowLab group applied deep-learning-based document classification models, including BlueBERT-Base [59], PubMedBERT [33], Joint Multilabel Attention Network (JMAN) [60], Hierarchical Label-wise Attention Networks (HLAN) [61], Hierarchical Attention Network Gated Recurrent Unit (HAN-GRU) [62], Hierarchical Attention Network (HAN) [63], CNN [38], Long short-term memory (LSTM) [64], etc., and each with a different combination of metadata (title, abstract, keywords and journal name), knowledge sources (UMLS, MeSH and Scientific Journal Rankings (SJR) journal categories), pretrained embeddings and data augmentation with back translation (to German). A class-specific ensemble averaging of the top-five models was then applied. The overall approach achieved micro-F1-scores of 0.9031 on the validation set and 0.8932 on the test set.

#### LIA/LS2N team [65]

We addressed the multi-label topic classification problem by combining an original keyword enhancement method with the TARS transformer-based approach [66] designed to perform few-shot learning. This model has first the advantage of not being constrained by the class number using a binary-like classification. Second, it tries to integrate the semantic information of the targeted class name in the training process by linking it to the content. Our best system architecture then uses a TARS model fed with various textual data sources such as abstracts, titles and keywords. Then, we applied a keyword-based enhancement that consists in applying a first-term frequency-inverse document frequency (TF-IDF) pass on the data to extract the specific terms of each topic with a score  $>0.65$ . These terms are then framed by tags [67], the idea being to explicitly give more importance to these terms during their modeling by the TARS model. Experiments conducted during the BioCreative challenge on the multi-label classification task show that our approach outperforms the baseline (ML-Net), no matter the metric considered, while being close to the best challenge approaches.

#### LRL\_NC team [68]

We propose two main techniques for this challenge task. The first technique is a data-centric approach, which uses insights on label co-occurrence patterns from the training data to segment the given problem into sub-problems. The second technique uses document-topic distribution extracted from contextual topic models as features for a binary relevance multi-label classifier. The best performance across different metrics was obtained using the first technique with TF-IDF representation of the raw text corpus as features. To solve each of these multi-label classification sub-problems, Random  $k$ -Labelsets classifier [69] was used with Light Gradient Boosting Machine (LGBM) [70] as base estimator.

### Opscidia team [71]

We propose creating an ensemble model by aggregating the sub-models at the end of each fine-tuning epoch with a weighting related to the Hamming loss. These models, based on BERT, are first pretrained on heterogeneous corpora in the scientific domain. The resulting meta-model is fed with several semi-independent samples augmented by random masking of COVID-19 terms, the addition of noise and the replacement of expressions with similar semantic features. While it is resource intensive if used directly, we consider its purpose to be distilling its rich new representation into a faster model.

### polyu\_cbsnlp team [72]

We propose an ensemble-learning-based method that utilizes multiple biomedical pretrained models. Specifically, we propose to ensemble seven advanced pretrained models for the LitCovid multi-label classification problem, including BioBERT-Base [32], BioBERT-Large [32], PubMedBERT [33], CovidBERT [73], BioELECTRA [57], BioM-ELECTRA [6] and BioMed\_RoBERTa [74], respectively. The homogeneous and heterogeneous neural architectures of these pre-training models assure the diversity and robustness of the proposed method. Furthermore, the extra biomedical knowledge of MeSH terms is also employed to enhance the semantic representations of the ensemble learning method. The final

**Table 6.** Overall team-submission-related statistics and the baseline performance. The baseline performance is the median of ten repetitions using different random seeds

	Label-based		Instance-based
	Macro-F1	Micro-F1	F1
Teams			
Mean	0.8191	0.8778	0.8931
Q1	0.7651	0.8541	0.8668
Median	0.8527	0.8925	0.9132
Q3	0.8670	0.9083	0.9254
Baseline			
ML-Net	0.7655	0.8437	0.8678

experimental results on the LitCovid shared task show the effectiveness and success of our proposed approach.

### robert-nlp team [75]

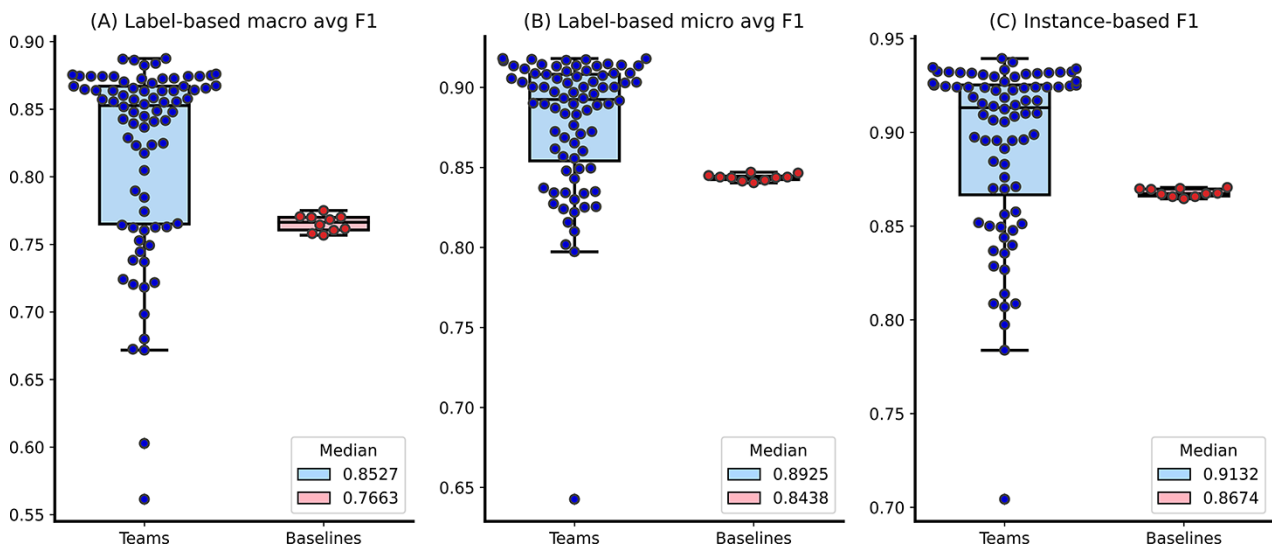
Our system represents documents using  $n$ -dimensional vectors using textual content (title and abstract) and metadata fields (pubtype, keywords and journal). Textual content and keywords are each encoded with SciBERT [48], and the two embeddings are concatenated. Following [76], this document representation is fed into a classification layer composed of several multi-layer perceptrons, each predicting the applicability of a single label. The model outperforms the shared task baseline both in terms of macro-F1 and in terms of micro-F1. Also, it is at par with the Q3 of the task statistics, which means that results are better than 75% of all the submitted runs.

### SINAI team [77]

To address the task of multi-label topic classification for COVID-19 literature annotation, the SINAI team opted for a problem transformation method that considers the prediction of each label as an independent binary classification task. This approach allowed the team to use the Logistic Regression algorithm [78] based on TF-IDF [79] representation of the tokenized and stemmed text data, which was previously subjected to a corpus augmentation process. This process consisted of using such techniques as back translation [80] of a selection of articles tagged with the less represented labels (Transmission, Case Report and Epidemic Forecasting) and the replacement of all nouns present in the abstracts with their synonyms retrieved from the WordNet [31]. The classifier achieved a label-based micro-average precision of 0.91, using negligible time and computational resources required to train our classifier addresses the fast growth of LitCovid.

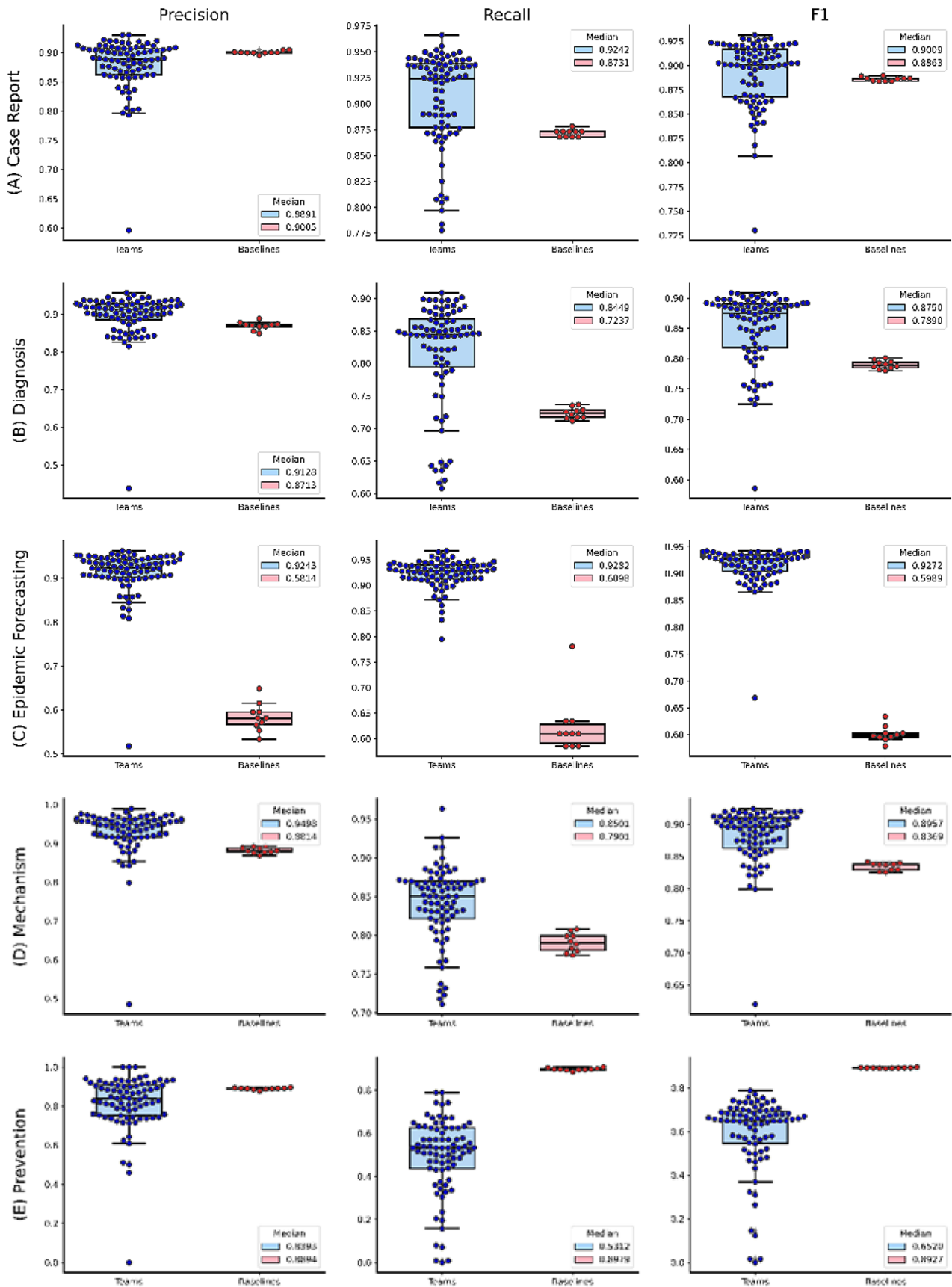
### TCS Research team [81]

We propose two different approaches for the task. The first approach, System 1, uses the training and validation datasets directly, whereas the second approach, System 2, performs



**Figure 2.** The distributions of team submission and baseline F1-scores. Median F1-scores are shown in the legend.





**Figure 3.** The distributions of team submission and baseline F1-scores for individual topics from (A) Case Report to (C) Epidemic Forecasting. Median F1-scores are shown in the legend. (B) The distributions of team submission and baseline F1-scores for individual topics from (D) Mechanism to (G) Treatment. Median F1-scores are shown in the legend.

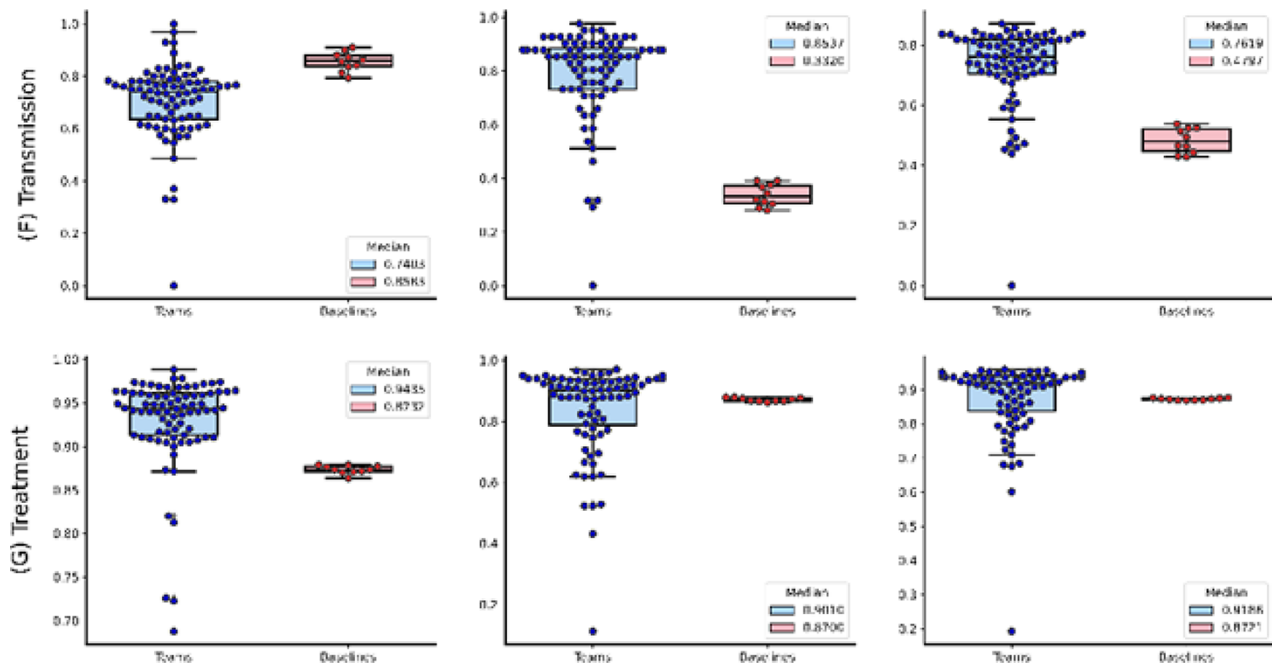


Figure 3. (Continued)

Table 7. Top-five team submission results ranked by each F1-score measure

Label-based		Instance-based	
Macro-F1	Micro-F1	F1	
Team	Result	Team	Result
Bioformer	0.8875	Bioformer	0.9181
Opscidia	0.8824	DUT914	0.9175
DUT914	0.8760	DonutNLP	0.9174
DonutNLP	0.8754	polyu_cbsnlp	0.9139
polyu_cbsnlp	0.8749	Opscidia	0.9135
		DUT914	0.9394
		DonutNLP	0.9346
		Bioformer	0.9334
		polyu_cbsnlp	0.9321
		ElsevierHealth Sciences	0.9307

named entity recognition (NER) on the training and validation datasets and uses the resulting tagged data for training/validation. NER on the abstract and title texts was performed using our text-mining framework PRIORI-T [82], where we cover 27 different entity types, including human genes, Severe Acute Respiratory Syndrome (SARS)/Middle East Respiratory Syndrome (MERS)/Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) genes, phenotypes, drugs, diseases, Gene Ontology (GO) terms, etc. In both approaches, training is performed by fine-tuning a BioBERT model pretrained on the Multi-Genre Natural Language Inference (MNLI) corpus [83]. Two separate BioBERT [32] fine-tuned models were created; the first model uses only the ‘abstract’ part of the training data and the second model uses only the ‘remaining’ part of the text, consisting of article title and metadata such as keywords and journal type. The final prediction was obtained by combining the predictions of both models, meaning that System 1 and System 2 each consist of a separate ensemble model. System 1 showed better performance than System 2 on both label- and instance-based F1-scores. Furthermore, System 1 showed better label-based macro- and instance-based F1-scores than the challenge baseline model (ML-Net) [14]. Finally, as per the challenge benchmarks, the label-based macro F1-score for System 1 was close

to the median F1-score and the instance-based F1-score was close to the mean score.

## Evaluation results

Table 6 summarizes team-submission-related statistics and the baseline performance in terms of their macro-F1-score, micro-F1-score and instance-based F1-score. The detailed results for each team submission and all the measures are provided in Table S1 in the supplementary material. The average macro-F1-score, micro-F1-scores and instance-based F1-scores are 0.8191, 0.8778 and 0.8931, respectively, all higher than the respective baseline scores. The baseline performance is close to the Q1 statistics for all the three measures, suggesting that ~75% of the team submissions have better performance than the baseline method.

Figure 2 shows the distributions of the overall performance, whereas Figure 3A and B further show the distributions of individual topic performance. Out of the seven topics, the teams achieved higher performance in terms of the median F1-score in six topics than the baseline (up to 29% higher) except the Prevention topic (only 4% lower). The results show that the performance difference is larger in the topics with relatively lower frequencies: Epidemic Forecasting

(23% higher) and Transmission (29% higher). In addition, we observe that the teams achieved generally consistent performance with the correlation of manual annotations in Table 3. For instance, it had the lowest performance on the Transmission topic, which is consistent with the correlation of manual annotations in Table 3. The only exception is the Epidemic Forecasting topic, where the inter-annotator agreement had a correlation of over 0.5, whereas the teams achieved an F1-score of over 0.9. This is primarily because of the sample size: only 5 and 41 articles are annotated with the Epidemic Forecasting topic in the random sample for inter-annotation agreement and the entire testing set, respectively. Given the limited size, we believe the performance on the Epidemic Forecasting topic is less representative. In contrast, other topics (which have a higher number of instances) show consistent performance.

Table 7 provides the top-five team submission performance ranked by each of the F1-scores. The best score is 6.8%, 4.1% and 4.1% higher than the corresponding team average score for macro-F1-score, micro-F1-score and instance-based F1-score, respectively. Four teams (Bioformer, DonutNLP, DUT914 and poly\_cbsnlp) consistently achieved top-ranked performance in the three rankings. As mentioned above, the Bioformer and DUT914 teams proposed innovative methods, which are beyond the default transfer learning approaches. In contrast, DonutNLP and poly\_cbsnlp used an ensemble of transformer approaches which also improve the performance. This is consistent with observations from previous challenge tasks [11, 24].

## Conclusions

This overview paper summarizes the BioCreative LitCovid track in terms of data collection and team participation. It provides a manually curated dataset of over 33 000 biomedical scientific articles. This is one of the largest datasets for multi-label classification for biomedical scientific literature, to our knowledge. Overall, 19 teams submitted 80 testing set predictions and ~75% of the submissions had better performance than the baseline approach. Given the scale of the dataset and the level of participation and team results, we conclude that the LitCovid track of BioCreative VII ran successfully and is expected to make significant contributions to innovative biomedical text-mining methods.

One possible direction to explore is the efficiency of transformers in real-world applications. As described above, over 80% of the teams used the transformers; the top-five team submissions also show superior performance using the transformer approach. However, it has a trade-off on the efficiency side. Existing studies show that transformers are significantly slower than other deep learning approaches using word and sentence embeddings, e.g. up to 80 times slower for biomedical sentence retrieval [84]. This is more challenging under the setting of multi-label classification (may require more than one transformer model) on COVID-19 literature (~10 000 articles per month). The Bioformer team showed one candidate approach, which only uses one-third of the parameters used by the original transformer architecture and achieves similar performance. We expect that more innovative transformer approaches will be developed to improve the efficiency.

Another possible direction is to quantify the usability of systems by incorporating them into the curation workflow.

The systems are ultimately used to facilitate data curation—it is thus important to evaluate its usability in the curation workflow, e.g. what is the accuracy of systems for new articles and how much manual curation effort can be reduced by deploying the systems? We have conducted a preliminary analysis on the generalization capability and efficiency of the systems in the LitCovid production environment [85], and we encourage more studies to perform usability evaluation and accountability of systems in the curation workflow [86, 87].

A further possible direction is the development of datasets for biomedical multi-label classification tasks. As summarized above, while multi-label classification is frequently used in biomedical literature, limited datasets are available for method development. This seems the major bottleneck for innovative biomedical text-mining methods. We expect that a community effort for dataset construction and a combination of automatic and manual curation approaches would address this issue. Also, given the scale of the BioCreative LitCovid dataset, it would be interesting to explore whether it can support transfer learning to other biomedical multi-label classification tasks. We encourage further development of biomedical text-mining methods using the BioCreative LitCovid dataset.

## Funding

Funding for open access charge: Intramural Research Program of the National Library of Medicine, National Institutes of Health.

## Conflict of interest

None declared.

## References

1. International Society for Biocuration (2018) Biocuration: distilling data into knowledge. *PLoS Biol.*, 16, e2002846. [10.1371/journal.pbio.2002846](https://doi.org/10.1371/journal.pbio.2002846).
2. Poux,S., Arighi,C.N., Magrane,M. *et al.* (2017) On expert curation and scalability: UniProtKB/Swiss-Prot as a case study. *Bioinformatics*, 33, 3454–3460. [10.1093/bioinformatics/btx439](https://doi.org/10.1093/bioinformatics/btx439).
3. Allot,A., Lee,K., Chen,Q. *et al.* (2021) LitSuggest: a web-based system for literature recommendation and curation using machine learning. *Nucleic Acids Res.*, 49. [10.1093/nar/gkab326](https://doi.org/10.1093/nar/gkab326).
4. Chen,Q., Leaman,R., Allot,A. *et al.* (2021) Artificial intelligence in action: addressing the COVID-19 pandemic with Natural Language Processing. *Annual Rev. Biomed. Data Sci.*, 4. [10.1146/annurev-biodatasci-021821-061045](https://doi.org/10.1146/annurev-biodatasci-021821-061045).
5. Chen,Q., Allot,A. and Lu,Z. (2020) Keep up with the latest coronavirus research. *Nature*, 579, 193. [10.1038/d41586-020-00694-1](https://doi.org/10.1038/d41586-020-00694-1).
6. Chen,Q., Allot,A. and Lu,Z. (2021) LitCovid: an open database of COVID-19 literature. *Nucleic Acids Res.*, 49, D1534–D1540. [10.1093/nar/gkaa952](https://doi.org/10.1093/nar/gkaa952).
7. Fabiano,N., Hallgrimson,Z., Kazi,S. *et al.* (2020) An analysis of COVID-19 article dissemination by Twitter compared to citation rates. *medRxiv*. [10.1101/2020.06.22.20137505](https://doi.org/10.1101/2020.06.22.20137505).
8. Yeganova,L., Islamaj,R., Chen,Q. *et al.* (2020) Navigating the landscape of COVID-19 research through literature analysis: a bird's eye view. preprint arXiv:2008.03397.
9. Ho,M.H.-C. and Liu,J.S. (2021) The swift knowledge development path of COVID-19 research: the first 150 days. *Scientometrics*, 126, 2391–2399. [10.1007/s11192-020-03835-5](https://doi.org/10.1007/s11192-020-03835-5).
10. Huang,-C.-C. and Lu,Z. (2016) Community challenges in biomedical text mining over 10 years: success, failure and the future. *Brief. Bioinformatics*, 17, 132–144. [10.1093/bib/bbv024](https://doi.org/10.1093/bib/bbv024).

11. Islamaj Doğan,R., Kim,S., Chatr-Aryamontri,A. *et al.* (2019) Overview of the BioCreative VI Precision Medicine Track: mining protein interactions and mutations for precision medicine. *Database*, 2019. [10.1093/database/bay147](https://doi.org/10.1093/database/bay147).
12. Arighi,C., Hirschman,L., Lemberger,T. *et al.* (2017) Bio-ID track overview. In: *Proceedings BioCreative Workshop*, 482, 376.
13. 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*. [10.48550/arXiv.2204.09781](https://doi.org/10.48550/arXiv.2204.09781).
14. Du,J., Chen,Q., Peng,Y. *et al.* (2019) ML-Net: multi-label classification of biomedical texts with deep neural networks. *J. Am. Med. Inform. Assoc.*, 26, 1279–1285. [10.1093/jamia/ocz085](https://doi.org/10.1093/jamia/ocz085).
15. Palayew,A., Norgaard,O., Safreed-Harmon,K. *et al.* (2020) Pandemic publishing poses a new COVID-19 challenge. *Nat. Hum. Behav.*, 4, 666–669. [10.1038/s41562-020-0911-0](https://doi.org/10.1038/s41562-020-0911-0).
16. Hanahan,D. and Weinberg,R.A. (2000) The hallmarks of cancer. *Cell*, 100, 57–70. [10.1016/S0092-8674\(00\)81683-9](https://doi.org/10.1016/S0092-8674(00)81683-9).
17. Larsson,K., Baker,S., Silins,I. *et al.* (2017) Text mining for improved exposure assessment. *PLoS One*, 12, e0173132. [10.1371/journal.pone.0173132](https://doi.org/10.1371/journal.pone.0173132).
18. Schober,P., Boer,C. and Schwarte,L.A. (2018) Correlation coefficients: appropriate use and interpretation. *Anesth. Analg.*, 126, 1763–1768. [10.1213/ANE.0000000000002864](https://doi.org/10.1213/ANE.0000000000002864).
19. Peters,M.E., Neumann,M., Iyyer,M. *et al.* (2018) Deep contextualized word representations. preprint arXiv:1802.05365.
20. Zhang,M.-L. and Zhou,Z.-H. (2013) A review on multi-label learning algorithms. *IEEE Trans. Knowl. Data Eng.*, 26, 1819–1837. [10.1109/TKDE.2013.39](https://doi.org/10.1109/TKDE.2013.39).
21. Nguyen,B.P. (2019) Prediction of FMN binding sites in electron transport chains based on 2-D CNN and PSSM profiles. *IEEE/ACM transactions on computational biology and bioinformatics*, 18, 2189–2197. [10.1109/TCBB.2019.2932416](https://doi.org/10.1109/TCBB.2019.2932416).
22. Le,N.Q.K., Do,D.T., Chiu,F.-Y. *et al.* (2020) XGBoost improves classification of MGMT promoter methylation status in IDH1 wildtype glioblastoma. *J. Pers. Med.*, 10, 128. [10.3390/jpm10030128](https://doi.org/10.3390/jpm10030128).
23. Wang,Y., Afzal,N., Liu,S. *et al.* (2018) Overview of the BioCreative/OHNLNLP challenge 2018 task 2: clinical semantic textual similarity In: *Proceedings of BioCreative/OHNLNLP Challenge*.
24. Chen,Q., Du,J., Kim,S. *et al.* (2018) Combining rich features and deep learning for finding similar sentences in electronic medical records. *Proceedings of BioCreative/OHNLNLP Challenge*. pp. 5–8.
25. Chen,Q., Panyam,N.C., Elangovan,A. *et al.* (2017) Document triage and relation extraction for protein-protein interactions affected by mutations. *Proceedings of the BioCreative VI Workshop*, 6, 52–51.
26. Madan,S., Szostak,J., Komandur Elayavilli,R. *et al.* (2019) The extraction of complex relationships and their conversion to biological expression language (BEL) overview of the BioCreative VI (2017) BEL track. *Database*, 2019. [10.1093/database/baz084](https://doi.org/10.1093/database/baz084).
27. Fang,L. and Wang,K. (2021) Team Bioformer at BioCreative VII LitCovid Track: Multi-label topic classification for COVID-19 literature with a compact BERT model. In: *Proceedings of the Seventh BioCreative Challenge Evaluation Workshop*.
28. Tang,W., Wang,J., Zhang,H. *et al.* (2021) Team DUT914 at BioCreative VII LitCovid Track: a BioBERT-based feature enhancement approach. In: *Proceedings of the Seventh BioCreative Challenge Evaluation Workshop*.
29. Bodenreider,O. (2004) The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Res.*, 32, D267–D270. [10.1093/nar/gkh061](https://doi.org/10.1093/nar/gkh061).
30. Wishart,D.S., Feunang,Y.D., Guo,A.C. *et al.* (2018) DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.*, 46, D1074–D1082. [10.1093/nar/gkx1037](https://doi.org/10.1093/nar/gkx1037).
31. Miller,G.A. (1995) WordNet: a lexical database for English. *Commun. ACM*, 38, 39–41. [10.1145/219717.219748](https://doi.org/10.1145/219717.219748).
32. Lee,J., Yoon,W., Kim,S. *et al.* (2020) BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36, 1234–1240. [10.48550/arXiv.1901.08746](https://doi.org/10.48550/arXiv.1901.08746).
33. Gu,Y., Tinn,R., Cheng,H. *et al.* (2020) Domain-specific language model pretraining for biomedical natural language processing. preprint arXiv:2007.15779.
34. Xu,S., Zhang,Y. and An,X. (2021) Team BJUT-BJFU at BioCreative VII LitCovid Track: a deep learning based method for multi-label topic classification in COVID-19 literature. In: *Proceedings of the Seventh BioCreative Challenge Evaluation Workshop*.
35. Sechidis,K., Tsoumakas,G. and Vlahavas,I. (2011) On the stratification of multi-label data. In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, Berlin, pp. 145–158.
36. Joulin,A., Grave,E., Bojanowski,P. *et al.* (2016) Fasttext. zip: compressing text classification models. preprint arXiv:1612.03651.
37. Lai,S., Xu,L., Liu,K. *et al.* (2015) Recurrent convolutional neural networks for text classification. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.
38. Kim,Y. (2014) Convolutional neural networks for sentence classification. *EMNLP*. *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
39. Vaswani,A., Shazeer,N., Parmar,N. *et al.* (2017) Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 6000–6010. [arXiv:1706.03762](https://arxiv.org/abs/1706.03762).
40. Xu,S. and An,X. (2019) ML2S-SVM: multi-label least-squares support vector machine classifiers. *Electron. Libr.* [10.1108/EL-09-2019-0207](https://doi.org/10.1108/EL-09-2019-0207).
41. Bagherzadeh,P. and Bergler,S. (2021) CLaC at BioCreative VII LitCovid Track: independent modules for multi-label classification of Covid articles. In: *Proceedings of the Seventh BioCreative Challenge Evaluation Workshop*.
42. Bagherzadeh,P. and Bergler,S. (2021) Multi-input recurrent independent mechanisms for leveraging knowledge sources: case studies on sentiment analysis and health text mining. *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pp. 108–118.
43. Alsentzer,E., Murphy,J.R., Boag,W. *et al.* (2019) Publicly available clinical BERT embeddings. preprint arXiv:1904.03323.
44. Lipscomb,C.E. (2000) Medical subject headings (MeSH). *Bull. Med. Libr. Assoc.*, 88, 265.
45. Bagherzadeh,P. and Bergler,S. (2021) Interacting knowledge sources, inspection and analysis: case-studies on biomedical text processing. In: *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*. pp. 447–456.
46. Bhatnagar,A., Bhavsar,N., Singh,M. *et al.* Team CUNI-NU at BioCreative VII LitCovid Track: multi-label topical classification of scientific articles using SPECTER embeddings with dual attention and label-wise attention network. In: *Proceedings of the Seventh BioCreative Challenge Evaluation Workshop*.
47. Cohan,A., Feldman,S., Beltagy,I. *et al.* (2020) Specter: document-level representation learning using citation-informed transformers. preprint arXiv:2004.07180.
48. Beltagy,I., Lo,K. and Cohan,A. (2019) SciBERT: A pretrained language model for scientific text. preprint arXiv:1903.10676.
49. Barbieri,F., Anke,L.E., Camacho-Collados,J. *et al.* Interpretable emoji prediction via label-wise attention LSTMs. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, New York, pp. 4766–4771.
50. Lin,S.-J., Chiu,Y.-W., Yeh,W.-C. *et al.* Team DonutNLP at BioCreativeVII LitCovid Track: multi-label topic classification for COVID-19 literature annotation using the BERT-based ensemble learning approach. In: *Proceedings of the Seventh BioCreative Challenge Evaluation Workshop*.

51. Tavchioski,I., Koloski,B., Škrlić,B. *et al.* Multi-label classification of COVID-19-related articles with an autoML approach. In: *Proceedings of the Seventh BioCreative Challenge Evaluation Workshop*.
52. Škrlić,B., Martinc,M., Lavrač,N. *et al.* (2021) autoBOT: evolving neuro-symbolic representations for explainable low resource text classification. *Mach. Learn.*, **110**, 989–1028. [10.1007/s10994-021-05968-x](https://doi.org/10.1007/s10994-021-05968-x).
53. Tian,S. and Zhang,J. Team FSU2021 at BioCreative VII LitCovid Track: BERT-based models using different strategies for topic annotation of COVID-19 literature. In: *Proceedings of the Seventh BioCreative Challenge Evaluation Workshop*.
54. Otmakhova,Y. and Yepes,A.J. Team ITTC at BioCreative VII LitCovid Track 5: combining pre-trained and bag-of-words models. In: *Proceedings of the Seventh BioCreative Challenge Evaluation Workshop*.
55. Zhang,T. (2004) Solving large scale linear prediction problems using stochastic gradient descent algorithms. In: *Proceedings of the Twenty-First International Conference on Machine Learning*, p. 116.
56. Yeganova,L., Comeau,D.C., Kim,W. *et al.* Text mining techniques for leveraging positively labeled data. In: *Proceedings of BioNLP 2011 Workshop*. pp. 155–163.
57. Raj Kanakarajan,K., Kundumani,B. and Sankarasubbu,M. (2021) BioELECTRA: pretrained biomedical text encoder using discriminators. In: *Proceedings of the 20th Workshop on Biomedical Language Processing*, pp. 143–154.
58. Dong,H., Wang,M., Zhang,H. *et al.* (2021) KnowLab at BioCreative VII Track 5 LitCovid: ensemble of deep learning models from diverse sources for COVID-19 literature classification. In: *Proceedings of the Seventh BioCreative Challenge Evaluation Workshop*, pp. 310–313.
59. Peng,Y., Yan,S. and Lu,Z. (2019) Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. preprint arXiv:1906.05474.
60. Dong,H., Wang,W., Huang,K. *et al.* (2020) Automated social text annotation with joint multilabel attention networks. *IEEE Trans. Neural Netw. Learn. Syst.*, **32**, 2224–2238. [10.1109/TNNLS.2020.3002798](https://doi.org/10.1109/TNNLS.2020.3002798).
61. Dong,H., Suárez-Paniagua,V., Whiteley,W. *et al.* (2021) Explainable automated coding of clinical notes using hierarchical label-wise attention networks and label embedding initialisation. *J. Biomed. Inform.*, **116**, 103728. [10.1016/j.jbi.2021.103728](https://doi.org/10.1016/j.jbi.2021.103728).
62. Yang,Z., Yang,D., Dyer,C. *et al.* (2016) 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*, pp. 1480–1489.
63. Baumel,T., Nassour-Kassis,J., Cohen,R. *et al.* (2018) Multi-label classification of patient notes: case study on ICD code assignment. In: *Workshops at the Thirty-Second AAAI Conference on Artificial Intelligence*.
64. Hochreiter,S. and Schmidhuber,J. (1997) Long short-term memory. *Neural Comput.*, **9**, 1735–1780. [10.1162/neco.1997.9.8.1735](https://doi.org/10.1162/neco.1997.9.8.1735).
65. Labrak,Y. and Dufour,R. (2021) Team LIA/LS2N at BioCreative VII LitCovid Track: multi-label document classification for COVID-19 literature using keyword based enhancement and few-shot learning. In: *Proceedings of the Seventh BioCreative Challenge Evaluation Workshop*.
66. Halder,K., Akbik,A., Krapac,J. *et al.* (2020) Task-aware representation of sentences for generic text classification. In: *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 3202–3213.
67. Caubrière,A., Rosset,S., Estève,Y. *et al.* (2020) Where are we in named entity recognition from speech? In: *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 4514–4520.
68. Tandon,K. and Chatterjee,N. (2021) LRL\_NC at BioCreative VII LitCovid Track: Multi-label classification of COVID-19 literature using ML-based approaches. In: *Proceedings of the Seventh BioCreative Challenge Evaluation Workshop*.
69. Tsoumakas,G., Katakis,I. and Vlahavas,I. (2010) Random k-labelsets for multilabel classification. *IEEE Trans. Knowl. Data Eng.*, **23**, 1079–1089. [10.1109/TKDE.2010.164](https://doi.org/10.1109/TKDE.2010.164).
70. Ke,G., Meng,Q., Finley,T. *et al.* (2017) Lightgbm: A highly efficient gradient boosting decision tree. *Adv Neural Inf Process Syst*, **30**, 3146–3154.
71. Rakotoson,L., Letailleur,C., Massip,S. *et al.* (2021) BagBERT: BERT-based bagging-stacking for multi-topic classification. preprint arXiv:2111.05808.
72. Gu,J., Wang,X., Chersoni,E. *et al.* (2021) Team polyU-CBSNLP at BioCreative-VII LitCovid Track: ensemble learning for COVID-19 multilabel classification. In: Vol. 24 of *Proceedings of the Seventh BioCreative Challenge Evaluation Workshop*, p. 2500.
73. Hebbar,S. and Xie,Y. (2021) CovidBERT-biomedical Relation Extraction for Covid-19. *Proceedings of FLAIRS-34*, **34**.
74. Alrowili,S. and Vijay-Shanker,K. BioM-Transformers: building large biomedical language models with BERT, ALBERT and ELECTRA. In: *Proceedings of the 20th Workshop on Biomedical Language Processing*, pp. 221–227.
75. Pujari,S.C., Tarsi,T., Strötgen,J. *et al.* Team RobertNLP at BioCreative VII LitCovid track: neural document classification using SciBERT. In: *Proceedings of the Seventh BioCreative Challenge Evaluation Workshop*.
76. Pujari,S.C., Friedrich,A. and Strötgen,J. A multi-task approach to neural multi-label hierarchical patent classification using transformers. In *European Conference on Information Retrieval*, pp. 513–528.
77. Chizhikova,M., López-Úbeda,P., Díaz-Galiano,M.C. *et al.* SINAI at BioCreative VII LitCovid Track: Corpus augmentation for COVID-19 literature multi-label classification. In: *Proceedings of the Seventh BioCreative Challenge Evaluation Workshop*.
78. Hilbe,J.M. (2009) *Logistic Regression Models*. Chapman and Hall/CRC, New York.
79. Salton,G. and Buckley,C. (1988) Term-weighting approaches in automatic text retrieval. *Inf Process Manag*, **24**, 513–523. [10.1016/0306-4573\(88\)90021-0](https://doi.org/10.1016/0306-4573(88)90021-0).
80. Junczys-Dowmunt,M., Grundkiewicz,R., Dwojak,T. *et al.* (2018) Marian: fast neural machine translation in C++. preprint arXiv:1804.00344.
81. Saipradeep,V., Sivadasan,N., Rao,A.R. *et al.* (2021) Team TCSR at BioCreative VII LitCovid Track: automated topic prediction of LitCovid using BioBERT. In: *Proceedings of the Seventh BioCreative Challenge Evaluation Workshop*.
82. Rao,A., Joseph,T., Saipradeep,V.G. *et al.* (2020) PRIORI-T: a tool for rare disease gene prioritization using MEDLINE. *PLoS One*, **15**, e0231728. [10.1371/journal.pone.0231728](https://doi.org/10.1371/journal.pone.0231728).
83. Williams,A., Nangia,N. and Bowman,S.R. (2017) A broad-coverage challenge corpus for sentence understanding through inference. preprint arXiv:1704.05426.
84. Chen,Q., Rankine,A., Peng,Y. *et al.* (2021) Benchmarking effectiveness and efficiency of deep learning models for semantic textual similarity in the clinical domain: validation study. *JMIR Med. Inform.*, **9**, e27386. [10.2196/27386](https://doi.org/10.2196/27386).
85. Chen,Q., Du,J., Allot,A. *et al.* (2022) LitMC-BERT: transformer-based multi-label classification of biomedical literature with an application on COVID-19 literature curation. preprint arXiv:2204.08649.
86. Wei,C.-H., Harris,B.R., Li,D. *et al.* (2012) Accelerating literature curation with text-mining tools: a case study of using PubTator to curate genes in PubMed abstracts. *Database*, **2012**. [10.1093/database/bas041](https://doi.org/10.1093/database/bas041).
87. Dowell,K.G., McAndrews-Hill,M.S., Hill,D.P. *et al.* (2009) Integrating text mining into the MGI biocuration workflow. *Database*, **2009**. [10.1093/database/bap019](https://doi.org/10.1093/database/bap019).