Overview of the BioCreative VII LitCovid Track: multi-label topic classification for COVID-19 literature annotation

Qingyu Chen[1], Alexis Allot[1], Robert Leaman[1], Rezarta Islamaj Doğan[1], Zhiyong Lu[1]
[1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA

***Abstract*—The BioCreative LitCovid track calls for a community effort to tackle automated topic annotation for COVID-19 literature. The number of COVID-19-related articles in the literature is growing by about 10,000 articles per month, significantly challenging curation efforts and downstream interpretation. LitCovid is a literature database of COVID-19-related articles in PubMed, which has accumulated more than 180,000 articles with millions of accesses each month by users worldwide. The rapid literature growth significantly increases the burden of LitCovid curation, especially for topic annotations. Topic annotation in LitCovid assigns one or more (up to eight) labels to articles. The annotated topics have been widely used both directly in LitCovid (e.g., accounting for ~20% of total uses) and downstream studies such as knowledge network generation and citation analysis. It is, therefore, important to develop innovative text mining methods to tackle the challenge.**

We organized the BioCreative LitCovid track to call for a community effort to tackle automated topic annotation for COVID-19 literature. This article summarizes the BioCreative LitCovid track in terms of data collection and team participation. The dataset is publicly available via https://ftp.ncbi.nlm.nih.gov/pub/lu/LitCovid/biocreative/. It consists of over 30K PubMed articles, one of the largest multi-label classification datasets on biomedical literature. There were 80 submissions in total from 19 teams worldwide. 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. We look forward to further participation in developing biomedical text mining methods in response to the rapid growth of the COVID-19 literature.

Keywords—biomedical text mining; natural language processing; artificial intelligence; machine learning; deep learning; multi-label classification; COVID-19; LitCovid;

I. 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 COVID-19 pandemic: the number of COVID-19-related articles in the literature is growing by about 10,000 articles per month. Figure 1 [4] shows that the median number of new COVID-19-related articles per day since May 2020 is 319, with a peak of over 2,500. This volume accounts for over 7% of all of PubMed.

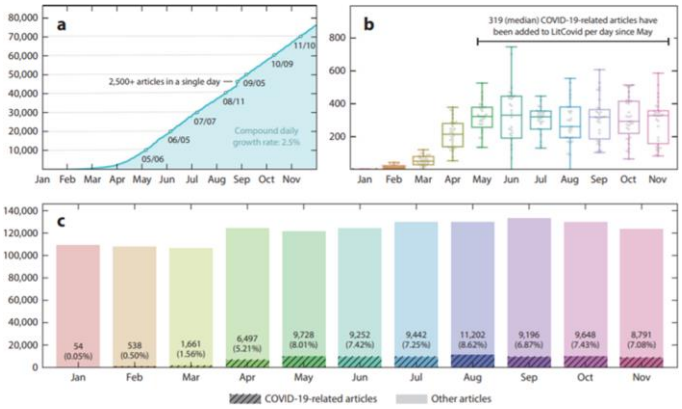

Figure 1. Growth of COVID-19-related articles in LitCovid. The figure is from the review article [4]. Figure 1a: the accumulative literature growth; Figure 1b: the number of COVID-19-related articles per day organized by months; Figure 1c: the ratio of COVID-19-related articles to the entire PubMed.

LitCovid [5, 6], a literature database of COVID-19-related articles in PubMed, has accumulated more than 180,000 articles, with millions of accesses each month by users worldwide. LitCovid is updated daily, and this rapid growth significantly increases 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. The annotated topics have been demonstrated to be effective for information retrieval and have been used in many downstream applications. Specifically, topic-related searching and browsing account for ~20% of LitCovid user behaviors [6], and the topics have also been used downstream studies such as citation analysis and knowledge network generation. application [7-9]. However, annotating these topics 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 nature of the task (assigning up to eight topics) and the ambiguity of natural languages (such as different ways to describe COVID-19 treatment procedures). While automatic approaches have been developed to assist manual curation, previous evaluations show that the automatic topic annotation tool has an F1-score of 10% lower than the tools assisting

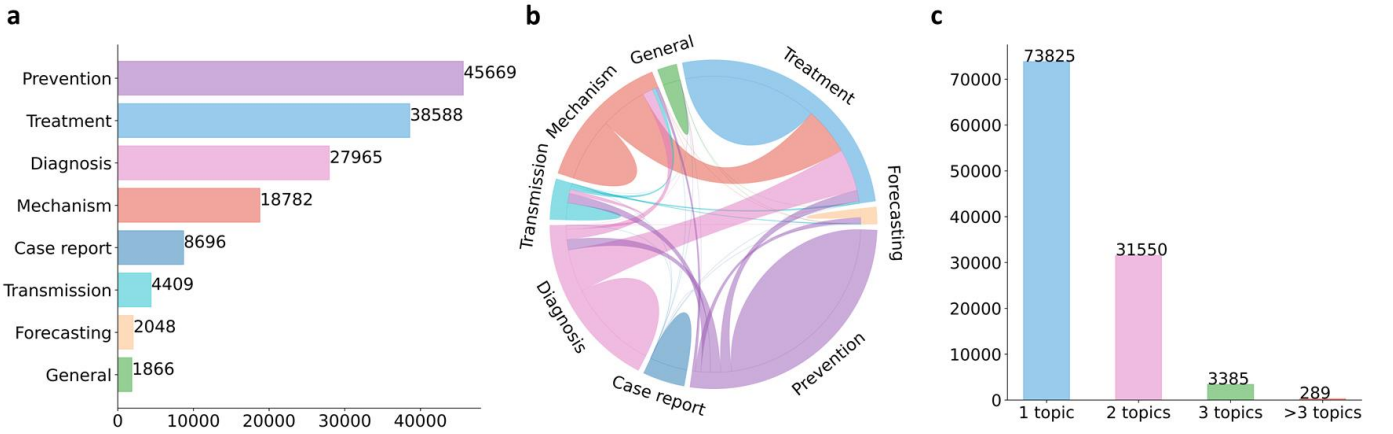

Figure 2. Characteristics of topic annotations in LitCovid up to September 2021. Figure 2a shows the frequencies of topics; Figure 2b demonstrates topic co-occurrences; and Figure 2c illustrates the distributions of the number of topics assigned per document.

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 and researchers worldwide.

We organized the BioCreative LitCovid track to call for a community effort to tackle automated topic annotation for COVID-19 literature. BioCreative is the first and longest-running community-wide effort for assessing biomedical text mining methods since 2003 [10]. Previous BioCreative challenges have successfully organized tracks on a range of biomedical text mining applications such as relation extractions [11] and entity normalizations [12].

This article provides an overview of the LitCovid track in terms of dataset collection and team participation. Overall, 19 teams submitted 80 runs and ~75% of the submissions had better performance than a shallow embedding baseline method [13]. 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, respectively. We encourage further work to develop biomedical text mining methods in response to the rapid growth of the COVID-19 literature.

II. LITCOVID CURATION PIPELINE

A. *The LitCovid curation pipeline*

The LitCovid curation pipeline is summarized in the LitCovid description article [6]. Essentially, the curation pipeline consists of three primary components: (1) document triage, identifying COVID-19 related articles from new articles in PubMed, (2) topic classification, assigning up to eight topics to the COVID-19 related articles (i.e., a multi-label classification task), and (3) entity recognition, extracting chemicals and locations mentioned in these articles. The curation pipeline has been performed daily since early February 2020.

Initially, the curation was done manually by two (part-time) human curators with a background in biomedical data sciences with little machine assistance. As the outbreak evolved, we developed automated approaches to support manual curation 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 F-1 scores were 0.99 and 0.94 for document triage and entity recognition, respectively). In contrast, the F-1 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.

B. *Topic classification in LitCovid*

Figure 2 shows the characteristics of annotated topics in LitCovid. The topic classification step assigns up to eight topics to the COVID-19 related articles, these are: (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, and (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, given their broad scope; e.g., a case report already contains diagnostic information by itself.

The 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 published COVID-19 articles without abstract information in PubMed are not descriptions of formal research studies but rather commentary or perspective [14]. 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 abstract available in PubMed, when the number of daily new articles reached a record high of over 2500.

III. DATASETS, BASELINE METHOD, AND EVALUATION MEASURES

A. *Training, development, and testing sets*

The training, development, and testing sets contain 24,960, 6,239, and 2,500 PubMed articles in LitCovid, respectively. 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 abstract available in PubMed.

The training and development datasets were made available June 15th to all participant teams. The testing set contains incoming hold-out articles added to LitCovid from 16th June to 22nd August. Using incoming articles to generate the testing set will facilitate the evaluation of the generalization capability of automatic tools.

B. *Baseline method*

We chose ML-Net [13] as our baseline method. ML-Net is a deep learning framework designed for multi-label classification tasks for biomedical literature. It combines contextual embeddings and the label counting mechanisms and achieved the best multi-label classification performance in the existing biomedical datasets.

C. *Evaluation measures*

We used both label-based and instance-based evaluation measures, the two most commonly used metrics for multi-label classification [15]. Label-based measures consider each label as an individual unit; the related measures calculate the performance for each label and then produce the aggregated measures for all the labels. In contrast, instance-based measures consider every individual instance as a unit. We calculated both macro and micro averages on Precision, Recall, and F1-score; we also calculated instance-based Precision, Recall, and F1-score summarized in the previous study [15].

Table 1. Team participation details sorted by team names alphabetically.

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
ElsevierHealthSciences	Elsevier	1
FSU2021	Florida State University	5
ittc	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	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

Table 2. Team submission-related statistics and the baseline performance.

	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

Table 3. Top 5 team submission results ranked by each F1-score measure.

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

IV. RESULTS AND DISCUSSIONS

A. Team submissions

Table 1 provides detailed participated teams and their number of submissions. Overall, 19 teams submitted 80 valid testing set predictions in total.

B. Overall performance and rankings

Table 2 summarizes team submission-related statistics and the baseline performance in terms of their macro F1-score, micro F1-score, and instance-based F1-score. We focus on the F1-scores because it aggregates both Precision and Recall. The detailed results for each team submission and all the measures are provided in Table S1. 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.

Furthermore, Table 3 provides top 5 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 average score for macro F1-score, micro F1-score, and instance-based F1-score, respectively. Four submissions (Bioformer, DonutNLP, DUT914, and polyu_cbsnlp) consistently achieved top-ranked performance in the three rankings.

V. CONCLUSION

This article summarizes the BioCreative LitCovid track in terms of its data collection and team participation. Overall, 19 teams submitted 80 testing set predictions and ~75% of the submissions had better performance than the shallow embedding baseline method. In the future, we plan to investigate the proposed methods, perform error analysis, and compare with the trade-offs of the evaluation metrics in more depth. We also encourage further development of biomedical text mining methods to respond to the rapid growth of the COVID-19 literature.

ACKNOWLEDGMENT

This work was supported by the Intramural Research Program of the National Library of Medicine, National Institutes of Health.

REFERENCES

1. International Society for Biocuration., *Biocuration: Distilling data into knowledge.* Plos Biology, 2018. **16**(4): p. e2002846.
2. Poux, S., et al., *On expert curation and scalability: UniProtKB/Swiss-Prot as a case study.* Bioinformatics, 2017. **33**(21): p. 3454-3460.
3. Allot, A., et al., *LitSuggest: a web-based system for literature recommendation and curation using machine learning.* Nucleic Acids Research, 2021.
4. Chen, Q., et al., *Artificial Intelligence in Action: Addressing the COVID-19 Pandemic with Natural Language Processing.* Annual Review of Biomedical Data Science, 2021. **4**.
5. Chen, Q., A. Allot, and Z. Lu, *Keep up with the latest coronavirus research.* Nature, 2020. **579**(7798): p. 193-193.
6. Chen, Q., A. Allot, and Z. Lu, *LitCovid: an open database of COVID-19 literature.* Nucleic Acids Research, 2021. **49**(D1): p. D1534-D1540.
7. Fabiano, N., et al., *An analysis of COVID-19 article dissemination by Twitter compared to citation rates.* medRxiv, 2020.
8. Yeganova, L., et al., *Navigating the landscape of COVID-19 research through literature analysis: a bird's eye view.* arXiv preprint arXiv:2008.03397, 2020.
9. Ho, M.H.-C. and J.S. Liu, *The swift knowledge development path of COVID-19 research: the first 150 days.* Scientometrics, 2021. **126**(3): p. 2391-2399.
10. Huang, C.-C. and Z. Lu, *Community challenges in biomedical text mining over 10 years: success, failure and the future.* Briefings in bioinformatics, 2016. **17**(1): p. 132-144.
11. Islamaj Doğan, R., et al., *Overview of the BioCreative VI Precision Medicine Track: mining protein interactions and mutations for precision medicine.* Database, 2019. **2019**.
12. Arighi, C., et al. *Bio-ID track overview*. in *Proc. BioCreative Workshop*. 2017.
13. Du, J., et al., *ML-Net: multi-label classification of biomedical texts with deep neural networks.* Journal of the American Medical Informatics Association, 2019. **26**(11): p. 1279-1285.
14. Palayew, A., et al., *Pandemic publishing poses a new COVID-19 challenge.* Nature Human Behaviour, 2020. **4**(7): p. 666-669.
15. Zhang, M.-L. and Z.-H. Zhou, *A review on multi-label learning algorithms.* IEEE transactions on knowledge and data engineering, 2013. **26**(8): p. 1819-1837.

Table S1. Detailed individual submission result.

	Label-based						Instance-based		
	Micro average			Macro average					
Team	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
Bioformer	0.9297	0.9038	0.9166	0.9038	0.8823	0.8875	0.9353	0.9269	0.9311
	0.9367	0.9002	0.9181	0.9120	0.8648	0.8839	0.9416	0.9234	0.9324
	0.9256	0.9085	0.9170	0.9032	0.8791	0.8863	0.9343	0.9285	0.9314
	0.9347	0.9018	0.9179	0.9191	0.8668	0.8870	0.9414	0.9256	0.9334
	0.9456	0.8650	0.9035	0.9184	0.8448	0.8743	0.9387	0.8960	0.9169
BJUT-BJFU	0.8474	0.8213	0.8342	0.8230	0.6744	0.7241	0.8508	0.8447	0.8477
	0.8775	0.8003	0.8371	0.8492	0.7107	0.7626	0.8490	0.8308	0.8398
	0.8555	0.8153	0.8349	0.8162	0.7050	0.7492	0.8404	0.8333	0.8368
	0.8704	0.8413	0.8556	0.8232	0.7596	0.7847	0.8731	0.8672	0.8701
	0.8982	0.8028	0.8478	0.8750	0.6913	0.7528	0.8670	0.8360	0.8512
CLaC	0.8766	0.8684	0.8725	0.8293	0.8281	0.8248	0.8958	0.8955	0.8956
	0.8762	0.8686	0.8724	0.8257	0.8316	0.8235	0.8955	0.8957	0.8956
	0.8810	0.8985	0.8896	0.8457	0.8598	0.8479	0.9023	0.9180	0.9101
	0.8808	0.8988	0.8897	0.8452	0.8633	0.8487	0.9024	0.9182	0.9102
CUNI-NU	0.8324	0.9007	0.8652	0.7472	0.8904	0.8047	0.8756	0.9232	0.8988
	0.9018	0.8764	0.8889	0.8554	0.8631	0.8570	0.9203	0.9046	0.9124
	0.9079	0.8673	0.8871	0.8617	0.8476	0.8478	0.9015	0.8898	0.8956
	0.9071	0.8205	0.8616	0.8938	0.8003	0.8409	0.8708	0.8448	0.8576
	0.9199	0.8731	0.8959	0.8824	0.8589	0.8673	0.9295	0.9016	0.9153
DonutNLP	0.9343	0.9010	0.9174	0.9214	0.8417	0.8725	0.9440	0.9254	0.9346
	0.9350	0.8946	0.9144	0.9152	0.8566	0.8754	0.9459	0.9222	0.9339
	0.9395	0.8852	0.9116	0.9047	0.8402	0.8646	0.9462	0.9137	0.9297
	0.9311	0.8963	0.9133	0.9084	0.8372	0.8639	0.9426	0.9223	0.9323
	0.9342	0.8877	0.9104	0.9157	0.8437	0.8702	0.9457	0.9190	0.9322
DUT914	0.9020	0.8985	0.9002	0.8567	0.8547	0.8447	0.9274	0.9270	0.9272
	0.9104	0.9190	0.9147	0.8778	0.8830	0.8760	0.9333	0.9417	0.9375
	0.9134	0.9217	0.9175	0.8791	0.8817	0.8744	0.9350	0.9438	0.9394
E8@IJS	0.8771	0.8114	0.8430	0.7611	0.6435	0.6799	0.8589	0.8449	0.8518
	0.8930	0.7826	0.8342	0.9175	0.6185	0.6724	0.8517	0.8200	0.8355
	0.8788	0.7757	0.8240	0.8720	0.6832	0.7382	0.8457	0.8121	0.8286
ElsevierHealthSciences	0.8979	0.9170	0.9074	0.8550	0.8892	0.8642	0.9244	0.9371	0.9307
FSU2021	0.9334	0.8841	0.9081	0.9204	0.8345	0.8670	0.9380	0.9117	0.9247
	0.9251	0.8814	0.9027	0.8979	0.8328	0.8577	0.9358	0.9095	0.9225
	0.9141	0.8803	0.8969	0.8917	0.8347	0.8518	0.9218	0.9062	0.9139
	0.9238	0.8880	0.9055	0.9035	0.8394	0.8635	0.9337	0.9143	0.9239
	0.9284	0.8861	0.9067	0.9062	0.8330	0.8636	0.9356	0.9131	0.9242
ittc	0.9210	0.8219	0.8686	0.8111	0.5664	0.6027	0.8715	0.8415	0.8562
	0.9136	0.8595	0.8857	0.8820	0.8398	0.8571	0.8976	0.8850	0.8913
	0.9242	0.8332	0.8764	0.9533	0.6318	0.6983	0.8874	0.8551	0.8710
	0.8861	0.9143	0.9000	0.8641	0.8764	0.8669	0.9058	0.9316	0.9185
KnowLab	0.8986	0.8850	0.8917	0.8965	0.8087	0.8416	0.9203	0.9135	0.9169

	0.9183	0.8637	0.8901	0.9049	0.8005	0.8426	0.9181	0.8951	0.9065
	0.9165	0.8711	0.8932	0.8990	0.8317	0.8601	0.9195	0.8977	0.9085
	0.9184	0.8626	0.8896	0.9078	0.7877	0.8288	0.9203	0.8914	0.9056
	0.9198	0.8501	0.8836	0.9124	0.7945	0.8393	0.9147	0.8807	0.8974
LIA/LS2N	0.5130	0.8598	0.6426	0.5240	0.7391	0.5614	0.5965	0.8597	0.7043
	0.8760	0.8659	0.8709	0.8498	0.8138	0.8231	0.8981	0.8942	0.8961
	0.8699	0.8966	0.8830	0.8298	0.8570	0.8366	0.8993	0.9198	0.9094
	0.8951	0.8280	0.8602	0.8814	0.7723	0.8174	0.8787	0.8610	0.8698
LRL_NC	0.8419	0.7572	0.7973	0.7645	0.6323	0.6717	0.8112	0.7840	0.7974
	0.8166	0.8844	0.8492	0.7652	0.7986	0.7624	0.8508	0.9028	0.8760
	0.8265	0.8894	0.8568	0.7781	0.8022	0.7742	0.8589	0.9085	0.8830
	0.8089	0.8465	0.8273	0.7543	0.7332	0.7217	0.8345	0.8660	0.8500
	0.8206	0.8473	0.8337	0.7840	0.7315	0.7445	0.8372	0.8622	0.8495
Opscidia	0.9302	0.8886	0.9089	0.9126	0.8567	0.8745	0.9363	0.9162	0.9261
	0.9325	0.8861	0.9087	0.9170	0.8401	0.8728	0.9359	0.9118	0.9237
	0.9310	0.8877	0.9088	0.9162	0.8432	0.8742	0.9351	0.9133	0.9241
	0.9369	0.8913	0.9135	0.9261	0.8544	0.8824	0.9409	0.9186	0.9296
	0.9309	0.8899	0.9099	0.9119	0.8551	0.8743	0.9366	0.9175	0.9270
PIDNA	0.9052	0.9004	0.9028	0.8853	0.8514	0.8633	0.9200	0.9246	0.9223
	0.9308	0.8816	0.9056	0.9131	0.8406	0.8635	0.9395	0.9099	0.9245
	0.9166	0.8844	0.9002	0.9111	0.8207	0.8583	0.9346	0.9138	0.9241
polyu_cbsnlp	0.8991	0.8874	0.8932	0.8840	0.8400	0.8551	0.9170	0.9120	0.9145
	0.9212	0.9057	0.9134	0.9139	0.8534	0.8749	0.9353	0.9279	0.9316
	0.9217	0.9049	0.9132	0.9016	0.8607	0.8742	0.9355	0.9281	0.9318
	0.9279	0.8999	0.9137	0.9078	0.8485	0.8692	0.9396	0.9243	0.9319
	0.9252	0.9029	0.9139	0.9099	0.8522	0.8726	0.9378	0.9264	0.9321
robert-nlp	0.9118	0.8888	0.9002	0.8566	0.8604	0.8555	0.9327	0.9176	0.9251
	0.9217	0.8855	0.9032	0.8897	0.8478	0.8655	0.9353	0.9132	0.9241
	0.8929	0.9015	0.8972	0.8465	0.8730	0.8536	0.9153	0.9224	0.9188
	0.9118	0.8888	0.9002	0.8566	0.8604	0.8555	0.9327	0.9176	0.9251
	0.9217	0.8855	0.9032	0.8897	0.8478	0.8655	0.9353	0.9132	0.9241
SINAI	0.8991	0.7370	0.8100	0.8343	0.7227	0.7629	0.8006	0.7676	0.7838
	0.9111	0.7486	0.8219	0.9004	0.6279	0.7203	0.8334	0.7820	0.8069
	0.9100	0.7553	0.8254	0.8840	0.6840	0.7603	0.8296	0.7887	0.8086
	0.9100	0.7547	0.8251	0.8829	0.6887	0.7643	0.8290	0.7891	0.8086
TCSR	0.8769	0.7624	0.8157	0.8542	0.6507	0.7181	0.8549	0.8003	0.8267
	0.8937	0.7746	0.8299	0.8542	0.7165	0.7653	0.8729	0.8167	0.8439
	0.8790	0.7370	0.8017	0.8350	0.6836	0.7370	0.8477	0.7825	0.8138
	0.8219	0.8791	0.8495	0.7721	0.8275	0.7896	0.8634	0.9066	0.8845