# LitCovid track Multi-label topic classification for COVID-19 literature annotation

### **README**

# Last updated on 12/09/2021

### Task description

The LitCovid track calls for a community effort to address automated topic annotation for COVID-19 literature. Topic annotation in LitCovid is a multi-label document classification task that assigns one or more labels to each article. There are 7 topic labels used in LitCovid: Treatment, Diagnosis, Prevention, Mechanism, Transmission, Epidemic Forecasting, and Case Report. These topics have been demonstrated to be effective for information retrieval and have also been used in many downstream applications related to COVID-19. However, annotating these topics manually has been a significant curation bottleneck. Increasing the accuracy of automated topic prediction in COVID-19-related literature would be a timely improvement beneficial to curators and researchers worldwide.

# Suggested references:

- a. For the track dataset:
  - Chen, Q., Allot, A., Leaman, R., Doğan, R.I. and Lu, Z., 2021. Overview of the BioCreative VII LitCovid Track: multi-label topic classification for COVID-19 literature annotation. In Proceedings of the seventh BioCreative challenge evaluation workshop.
- b. For the LitCovid in general:
  - Chen, Q., Allot, A. and Lu, Z., 2020. Keep up with the latest coronavirus research. Nature, 579(7798), pp.193-193.
  - Chen, Q., Allot, A. and Lu, Z., 2021. <u>LitCovid: an open database of COVID-19</u> literature. Nucleic Acids Research, 49(D1), pp. D1534-D1540.
- c. For the baseline method:
  - Du, J., Chen, Q., Peng, Y., Xiang, Y., Tao, C. and Lu, Z., 2019. ML-Net: multi-label classification of biomedical texts with deep neural networks. Journal of the American Medical Informatics Association, 26(11), pp.1279-1285. References

# Datasets (updated in 12/09/2021)

The topics of the articles in all the datasets have been manually reviewed:

- a. BC7-LitCovid-Train.csv: the training set contains 24,960 articles from LitCovid;
- b. BC7-LitCovid-Dev.csv: the validation set contains 6,239 articles from LitCovid;
- c. BC7-LitCovid-Test.csv: the test set contains 2,500 articles from LitCovid, released during the challenge;
- d. BC7-LitCovid-Test-GS.csv: the test set contains 2,500 articles from LitCovid with gold standard labels released after the challenge.

#### File format

The datasets are provided in csv format, with the following fields retrieved from PubMed/LitCovid:

- pmid: PubMed Identifier
- journal: journal name
- title: article title
- abstract: article abstract
- keywords: author-provided keywords
- pub\_type: article type, e.g., journal article
- authors: author names

- doi: Digital Object Identifier
- label: annotated topics, i.e., the output
  - each article can be assigned one or more labels (Treatment, Diagnosis, Prevention, Mechanism, Transmission, Epidemic Forecasting, and Case Report)
  - each label is separated by a semicolon, e.g., 'Diagnosis; Treatment' means that the article is assigned both the label Diagnosis and the label Treatment

Other fields are also available via <a href="https://ftp.ncbi.nlm.nih.gov/pub/lu/LitCovid/">https://ftp.ncbi.nlm.nih.gov/pub/lu/LitCovid/</a>, which provides additional fields if needed, such as biological entity annotations.

# Evaluation script (updated in 08/31/21)

Submissions will be evaluated using both label-based and instance-based metrics that are commonly applied for multi-label classification. Evaluation scripts are provided via <a href="https://github.com/ncbi/biocreative\_litcovid">https://github.com/ncbi/biocreative\_litcovid</a>. An example of prediction file is also provided via <a href="https://github.com/ncbi/biocreative\_litcovid/blob/main/prediction\_label\_samples.csv">https://github.com/ncbi/biocreative\_litcovid/blob/main/prediction\_label\_samples.csv</a>. The submission instruction will be available later.

# Submission instructions (updated in 08/31/21)

- You may submit up to five runs (predictions).
- Submissions are due 12<sup>th</sup> September 11:59:59 "Anywhere on Earth".
- The prediction files must follow the same format as
  <a href="https://github.com/ncbi/biocreative\_litcovid/blob/main/prediction\_label\_samples.csv">https://github.com/ncbi/biocreative\_litcovid/blob/main/prediction\_label\_samples.csv</a> and the evaluation script in <a href="https://github.com/ncbi/biocreative\_litcovid">https://github.com/ncbi/biocreative\_litcovid</a> also validates the file format.
- Please submit each prediction separately with specific naming format (see the steps below).
- Submission steps:
  - 1. Please go to https://easychair.org/conferences/?conf=bc7
  - 2. Select 'Track 5- LitCovid track Multi-label topic classification for COVID-19 literature annotation' and click 'Continue', which will lead to the submission page
  - 3. In the submission page:
    - Please fill in the details of team members (authors) as requested
    - In the title field, please follow the format of team\_name-submission\_id, e.g., WorldPeaceTeam Submission1
    - In the abstract field, please provide a brief summary of the methods and models for this specific submission (up to one paragraph)
    - Please provide keywords as requested
    - File upload:
      - Please place your submission file into a folder and zip it (the submission portal requires the submission file type as zip)
      - Optionally, you could provide additional readme files into the folder if needed
      - Upload the zipped folder
    - Click 'Submit'
    - Repeat the steps for another submission (up to five)

# Contract

Please contact <a href="mailto:qio.guestions">qingyu.chen@nih.gov</a> with the subject heading "BioCreative Track 5 LitCovid questions" if you have any questions.

### Status updates and FAQs

We will also provide updates and FAQs via

 $\underline{https://biocreative.bioinformatics.udel.edu/tasks/biocreative-vii/track-5/}$