Data statement for ‘CORD-19-Vaccination’

Manisha Singh, Divy Sharma, Alonso Ma, Bridget Tyree

Contents

1. HEADER

Dataset Title

The title of the dataset is 'CORD-19-Vaccination': 'COVID-19 Open Research Dataset - Vaccination'.

Dataset Curator(s) [name, affiliation]

Name: Manisha Singh(manishas@uw.edu), Divy Sharma(divy@uw.edu),Alonso Ma(amatake@uw.edu), and Bridget Tyree (btyree@uw.edu)
Affiliation: University of Washington

Dataset Version [version, date]

Latest release : [V1 (CORD-19-vacination_2022-06-06.tar.gz), 2022-06-06]
Further note on Version and dates:
CORD-19-Vaccination is based on the CORD-19 version [v1, 2022-05-03]
CORD-19 Releases: https://ai2-semanticscholar-cord-19.s3-us-west-2.amazonaws.com/historical_releases.html

Dataset Citation and DOI:

Dataset Citation:
Singh M., Sharma D., Ma A., Tyree B. (June 2022). CORD-19-Vaccination: The COVID-19 Open Research Dataset Vaccination Subset. Unpublished manuscript. University of Washington (UW)

Original Dataset Citation:

Wang, L.L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Burdick, D., Eide, D., Funk, K., Katsis, Y., Kinney, R.M., Li, Y., Liu, Z., Merrill, W., Mooney, P., Murdick, D.A., Rishi, D., Sheehan, J., Shen, Z., Stilson, B., Wade, A.D., Wang, K., Wang, N.X.R., Wilhelm, C., Xie, B., Raymond, D.M., Weld, D.S., Etzioni, O. & Kohlmeier, S. (July 2020). CORD-19: The COVID-19 Open Research Dataset. *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*. Online: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/2020.nlpcovid19-acl.1

Original Dataset DOI:

https://doi.org/10.48550/arXiv.2004.10706

Data Statement Author(s) [name, affiliation]

[Manisha Singh(manishas@uw.edu), Divy Sharma(divy@uw.edu), Alonso Ma(amatake@uw.edu) and Bridget Tyree (btyree@uw.edu), University of Washington]

Data Statement Version [version, date]

[v1, 2022-06-06]

Data Statement Citation

Singh M., Sharma D., Ma A., Tyree B. (June 2022). Data statement for ‘CORD-19-Vaccination’. Version Unpublished manuscript. University of Washington (UW)

2. EXECUTIVE SUMMARY

'CORD-19-Vaccination' is a corpus of academic papers on COVID-19 vaccination and related research. ‘CORD-19-Vaccination’ dataset is a resource dataset to support text mining and NLP research in the domain of COVID-19 vaccination. This dataset consists of the metadata of all the papers published in the space of COVID-19 vaccination in all the languages worldwide. This corpus is intended for use by people who want to investigate COVID-19 vaccine related research. This research took place between 2020 and 2022. The dataset comprises of a metadata.csv which consist of approximately 30K rows. Figure 1 gives an overall representation of the CORD-19-Vaccination dataset.

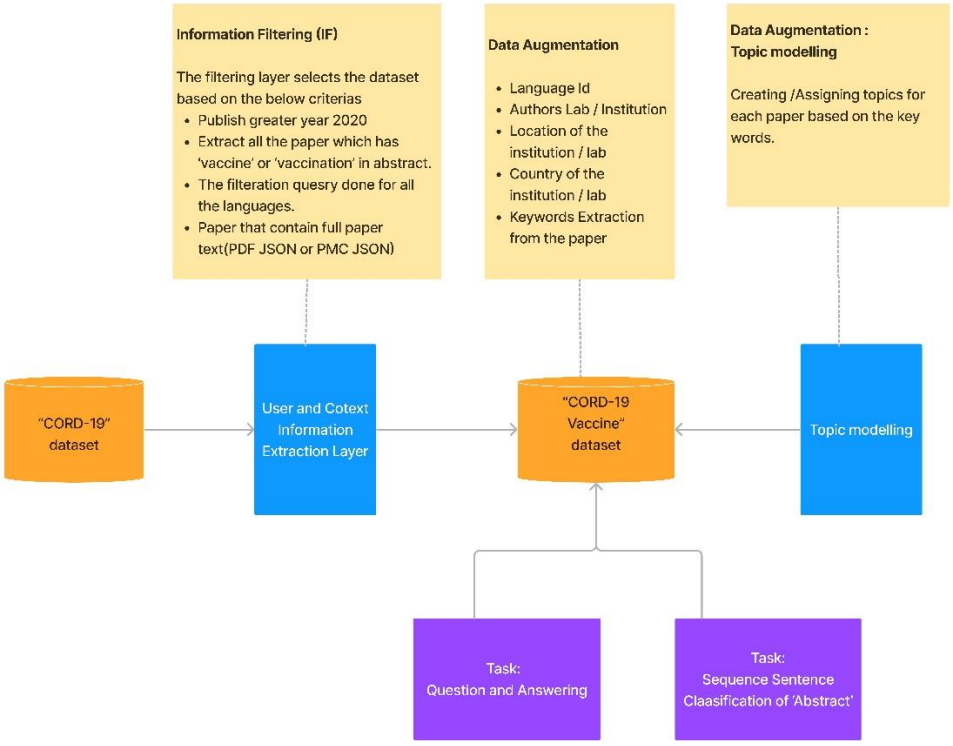


Figure 1 :CORD-19-Vaccination dataset - Overview

3. CURATION RATIONALE

'CORD-19-Vaccination' is a resource dataset, extracted from the CORD-19 dataset that can be used by the researchers for vaccination related research. The dataset will solve the information overload issue of CORD-19 dataset for those specifically looking into 'vaccine' research. Information Filtering (IF) was done to extract only the relevant information relating to 'vaccine' from CORD-19. The main filtering criteria were 'publish date' and 'pattern search (vaccine)'.

The CORD-19-Vaccination dataset was augmented by adding the following fields: "language ID", "affiliation", "country", and "keywords used in the text of the paper". These fields were not part of the CORD-19 dataset. These additional features for each research paper enable the user to get the details of the paper quickly. This enhances the filtering and selection process for research papers and improves the overall user experience.

The availability of the data "language ID" will enable the downstream task to classify, retrieve or train an application based on the language of the text. The availability of "affiliation" and "country" will help downstream tasks including collaborations between institutions and countries. The fields "language Id", "affiliation" and "country" will help the user to identify the text demography of the whole dataset. The availability of the data "keywords used in the text of the paper" will help in information extraction tasks based on keyword search.

Topic modelling will provide a comprehensive view of the dataset's main contents, allowing potential users to quickly assess common themes among the papers and have an overview of the dataset's distribution. The generated topics will also enable the users to further curate the dataset according to specific needs/context of the downstream task they are implementing. For example, when implementing a Question Answering task to respond queries regarding COVID-19 vaccine's side effects, one might choose to subset the training dataset to the corresponding topic.

Task: We have implemented two tasks ‘Question and answering’ and ‘Sentence Sequence classification’ for CORD-19-Vaccination dataset. The questions are relevant to COVID- 19 vaccine and like Kaggle competition task for CORD-19 dataset. Sentence sequence classification of the abstract of each paper into the specific sections "Background/Objective", "Method", "Result/Conclusion" will help users to access the research paper’s overall idea and make the abstract easy to read. At the same time this will enable downstream tasks for ‘Abstract’ classification. “Automatically classifying each sentence in an ‘Abstract’ will help re-searchers read abstracts more efficiently, especially in fields where abstracts may belong, such as the medical field” (Dernoncourt, Lee, & Szolovits 2016).

3.1 Data collection logic

The overall goal was to create a dataset that was based out of CORD-19 but only included the papers that are relevant to vaccine research. The work was done in three phases.

Extraction phase: The first phase was the extraction phase. In this phase we created the data pipeline . Then using SQLite query we created a subset of the dataset from CORD-19 dataset, and took only those papers where the starting ‘publish time’ was ‘2020’ and either the ‘Abstract’ or ‘Title’ contains the word ‘vaccine’ in any language.

Data augmentation phase: The second phase is the data augmentation phase. In this phase we added some more columns to the dataset. The columns were added based on tasks such as ‘Question and Answer’ and ‘Topic modeling’ and other such NLP related tasks. The data ‘author lab/institution’ and ‘lab/institution country’ was mainly collected from the ‘json parse’ files of the research papers. We automatically Google searched each research paper to get the ‘lab/institution country’. The language was added using Facebook AI ‘fastText’, and keyword was added using ‘Yake’. Finally, we implemented ‘Topic modelling’ where we classified the dataset into topics based on the ‘abstract’ using the LDA model. Each implementation is detailed below in its respective section.

Task implementation phase: We implemented the ‘Question and Answering’ task and Sequential Sentence Classification task using the CORD-19-Vaccination dataset. The result and implementation details are in section 9.3.

3.2 Internal organization of the dataset and data constituents

The dataset consist of the metadata.csv which consists of the fields below. Most of the columns directly come from the CORD-19 dataset, and have the same column description as mentioned in the data statement of CORD-19 attached in the Appendix. The columns in ‘blue’ were extracted from CORD-19. The columns in ‘purple’ are from the CORD -19-Vaccination dataset added as part of the data augmentation.

Column Id	Column description	data type	Example values

cord_uid	A str-valued field that assigns a unique identifier to each CORD-19 paper. This is not necessarily unique per row, which is explained in the FAQs	string	*d1pd09zj*
sha	A List[str]-valued field that is the SHA1 of all PDFs associated with the CORD-19 paper. Most papers will have either zero or one value here (since either have a PDF or we don't), but some papers will have multiple. For example, the main paper might have supplemental information saved in a separate PDF. Or might have two separate PDF copies of the same paper. If multiple PDFs exist, their SHA1 will be semicolon-separated (e.g. '4eb6e165ee705e2ae2a24ed2d4e67da42831ff4a; d4f0247db5e916c20eae3f6d772e8572eb828236')	string	*1cee4a0d0e823379ec34a462a04561bf4cd736a2*
source_x	A List[str]-valued field that is the names of sources from which we received this paper. Also semicolon-separated. For example, 'ArXiv; Elsevier; PMC; WHO'. There should always be at least one source listed	string	*PMC*
title	A str-valued field for the paper title	string	*Synthetic carbohydrate-based vaccines: challenges and opportunities*
doi	A str-valued field for the paper DOI	string	*10.1186/s12929-019-0591-0*
pmcid	A str-valued field for the paper's ID on PubMed Central. Should begin with PMC followed by an integer.	string	*PMC6941340*
pubmed_id	An int-valued field for the paper's ID on PubMed.	integer	*31900143*
license	A str-valued field with the most permissive license we've found associated with this paper. Possible values include: 'cc0', 'hybrid-oa', 'els-covid', 'no-cc', 'cc-by-nc-sa', 'cc-by', 'gold-oa', 'biorxiv', 'green-oa', 'bronze-oa', 'cc-by-nc', 'medrxiv', 'cc-by-nd', 'arxiv', 'unk', 'cc-by-sa', 'cc-by-nc-nd'	string	*cc-by*
abstract	A str-valued field for the paper's abstract	string	*Glycoconjugate vaccines based on bacterial capsular polysaccharides (CPS) have been extremely successful in preventing bacterial infections. The glycan antigens for the preparation of CPS based glycoconjugate vaccines are mainly obtained from bacterial fermentation, the quality and length of glycans are always inconsistent. Such kind of situation make the CMC of glycoconjugate vaccines are difficult to well control. Thanks to the advantage of synthetic methods for carbohydrates syntheses. The well controlled glycan antigens are more easily to obtain, and them are conjugated to carrier protein to from the so-call homogeneous fully synthetic glycoconjugate vaccines. Several fully glycoconjugate vaccines are in different phases of clinical trial for bacteria or cancers. The review will introduce the recent development of fully synthetic glycoconjugate vaccine.*
publish_time	A str-valued field for the published date of the paper. This is in yyyy-mm-dd format. Not always accurate as some publishers will denote unknown dates with future dates like yyyy-12-31	string	*1/3/2020*
authors	A List[str]-valued field for the authors of the paper. Each author name is in Last, First Middle format and semicolon-separated.	string	*Mettu, Ravinder; Chen, Chiang-Yun; Wu, Chung-Yi*
journal	A str-valued field for the paper journal. Strings are not normalized (e.g. BMJ and British Medical Journal can both exist). Empty string if unknown.	string	*J Biomed Sci*
mag_id	Deprecated, but originally an int-valued field for the paper as represented in the Microsoft Academic Graph.	integer	

who_covidence_id	A str-valued field for the ID assigned by the WHO for this paper. Format looks like #72306.	string	
arxiv_id	A str-valued field for the arXiv ID of this paper.	string	
pdf_json_files	A List[str]-valued field containing paths from the root of the current data dump version to the parses of the paper PDFs into JSON format. Multiple paths are semicolon-separated. Example: document_parses/pdf_json/4eb6e165ee705e2ae2a24ed2d4e67da42831ff4a.json; document_parses/pdf_json/d4f0247db5e916c20eae3f6d772e8572eb828236.json	string	*document_parses/pdf_json/1cee4a0d0e823379ec34a462a04561bf4cd736a2.json*
pmc_json_files	A List[str]-valued field. Same as above, but corresponding to the full text XML files downloaded from PMC, parsed into the same JSON format as above	string	*document_parses/pmc_json/PMC6941340.xml.json*
url	A List[str]-valued field containing all URLs associated with this paper. Semicolon-separated.	string	*https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6941340/*
s2_id	A str-valued field containing the Semantic Scholar ID for this paper. Can be used with the Semantic Scholar API (e.g. s2_id=9445722 corresponds to http://api.semanticscholar.org/corpusid:9445722	string	
lang_id	Language identifier: The language id for which the fastText model's confidence is highest becomes the language id for that paper.	String	*en*
lang_id_confidence	Language id confidence: This is score assigned to the language id by the fastText model.	String	*0.9167*
lang_id_predictions	Language id predictions: This column gives the top three scores given by the fastText model.	string	*en=0.9167, id=0.0055, fr=0.0043*
aff_lab_inst	Affiliation Lab/Institution: This field gives the first author's Lab/Institution for each paper.	String	*University of Maryland School of Medicine*
Aff_location	Affiliation location: The location of the lab/institution.	string	*postCode=21201; region=MD; settlement=Baltimore*
Aff_country	Affiliation country: The country of the lab/institution.	String	*USA*
Keywords	Keywords: Extracted from the 'Title', 'Abstract' and body of the text using Yake. The list contains the top 20 keywords.	String	*DNA vaccine; archaeosome; DNA; recombinant DNA vaccine; pDNA - surface localized archaeosome ; archaeosome vaccines group; cells; DNA vaccine candidate; localized archaeosome; vaccine; archaeosome vaccines; groups; plasmid DNA; gene DNA vaccine; PBS control groups; recombinant gene; pDNA-encapsulated archaeosomes; gene; mice; control groups*
Labelled_abstract	Labeled Abstract: The Abstract is passed through sequential sentence classification and the result is every sentence of the abstract is labelled with one of the following labels: 'Background, Objective, Method, Result, Conclusion '	String	*BACKGROUND: Glycoconjugate vaccines based on bacterial capsular polysaccharides (CPS) have been extremely successful in preventing bacterial infections. BACKGROUND: The glycan antigens for the preparation of CPS based glycoconjugate vaccines are mainly obtained from bacterial fermentation, the quality and length of glycans are always inconsistent.*

			BACKGROUND: Such kind of situation make the CMC of glycoconjugate vaccines are difficult to well control. CONCLUSIONS: Thanks to the advantage of synthetic methods for carbohydrates syntheses.
topic	Topic: Label for the inferred topic of the paper. The label corresponds to the topic which had the highest probability among the predicted topic distribution.	String	*Vaccine development; Vaccination side-effects / Treatments*
topic_index	Topic Index: Index associated with the inferred topic for the paper, can take values between 0 and 4.	integer	*0; 1; 2; 3; 4*
topic_prob	Topic Probability: Probability of the given paper corresponding to the assigned topic label.	float	*0.524614*
std_first_auth_country	Standardized First Author Country: Name of the country affiliated with the first author of the paper, standardized to match the Geopandas ''naturalearth_lowres'' country names.	string	*Taiwan; United States of America*

4. DOCUMENTATION FOR SOURCE DATASETS

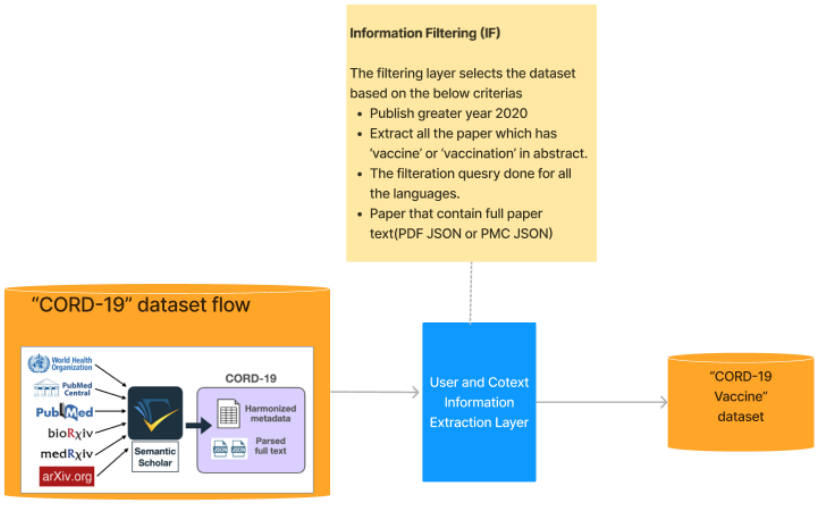


Figure 2: Data Source flow for CORD-19-Vaccination

The "CORD-19-Vacination" dataset was extracted from the "CORD-19" dataset. As you can see in the Figure 2, papers in the CORD-19 dataset are sourced from PubMed Central (PMC), PubMed, the World Health Organization's Covid-19 Database, and preprint servers bioRxiv, medRxiv, and arXiv. The source of all CORD-19 papers is in the introduction section of the paper (Wang et al. 2020). *Semantic Scholar API*: Metadata, paper abstracts, and citation information for papers we index are available through our API. Documentation here (Semantic Scholar 2022). Publishers such as Elsevier (here) and Springer Nature (here) provide full text coverage of relevant papers available in their back catalog (Springer Nature 2022; Elsevier 2022).

After extracting all papers from CORD-19 which were published after 2020 and had the word *vaccine* in either the abstract or the title, we augmented the CORD-19-Vaccination dataset by adding the following columns: ‘Language ID’, ‘Author lab/institution affiliation’, ‘Keywords from Abstract/Title/Body text’, ‘Topic’ from topic modeling, and ‘Abstract classification’ from sequential sentence classification.

5. LANGUAGE VARIETIES

Figure 3 below shows the number of articles/papers/journals in each language which was gathered from our Exploratory Data Analysis (EDA) of the CORD-19 dataset.

Figure 3: Language distribution in CORD-19

Since the ‘CORD-19-Vaccination’ dataset was extracted from the ‘CORD-19’ dataset, in order to capture the word 'vaccine/vaccination' in every language the query for the information extraction was customized to search the pattern of 'vaccine/vaccination' in every language. A small snippet of the code is shown in Figure 4. For the full code please look at the Information Filtering code here.

```
title LIKE '%مصل%' OR abstract LIKE '%مصل%' OR
title LIKE '%peyvənd%' OR abstract LIKE '%peyvənd%' OR
title LIKE '%ваксина%' OR abstract LIKE '%ваксина%' OR
title LIKE '%vacuna%' OR abstract LIKE '%vacuna%' OR
title LIKE '%vakcína%' OR abstract LIKE '%vakcína%' OR
title LIKE '%vaccine%' OR abstract LIKE '%vaccine%' OR
title LIKE '%Impfung%' OR abstract LIKE '%Impfung%' OR
title LIKE '%εμβόλιο%' OR abstract LIKE '%εμβόλιο%' OR
title LIKE '%vaccine%' OR abstract LIKE '%vaccine%' OR
title LIKE '%vacuna%' OR abstract LIKE '%vacuna%' OR
```

Figure 4: Snippet of the 'vaccine/vaccination' pattern search in different languages

The result of this is that the ‘CORD-19- Vaccination’ dataset consists of papers in ‘English’, ‘French’, ‘Spanish’, ‘Dutch’, ‘German’, ‘Portuguese’ and ‘Italian’ as shown in Figure 5.

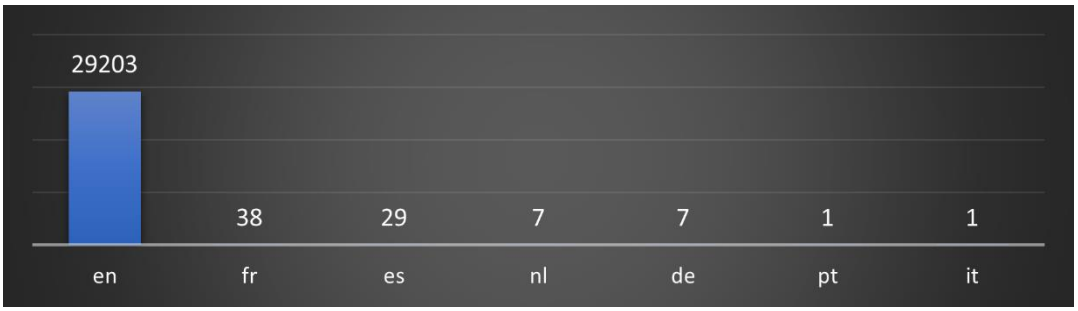


Figure 5: Language distribution CORD-19-Vaccination

6. SPEAKER DEMOGRAPHIC

Speaker/Author demographic was mainly assessed via examination of the distribution of first author's countries of affiliation. Country of affiliation was initially extracted from the full text JSON files with coverage of around 63% (~18.5K) of the total of papers. Through web scraping (see Section 9.2.1.2), country of affiliation was identified for an additional group of papers, increasing coverage to 93% (~27K).

50% of the total papers are concentrated over 7 countries: United States of America, China, India, Italy, United Kingdom, Germany and Canada, with the USA representing 20% of the dataset. A complete map depicting the distribution of number of papers by country of affiliation of the first author can be observed in Figure 6. Most notably, apart from the concentration of research in the previously mentioned countries, a stark lack of representation from the Global South (with the exception of Brazil) is also evident.

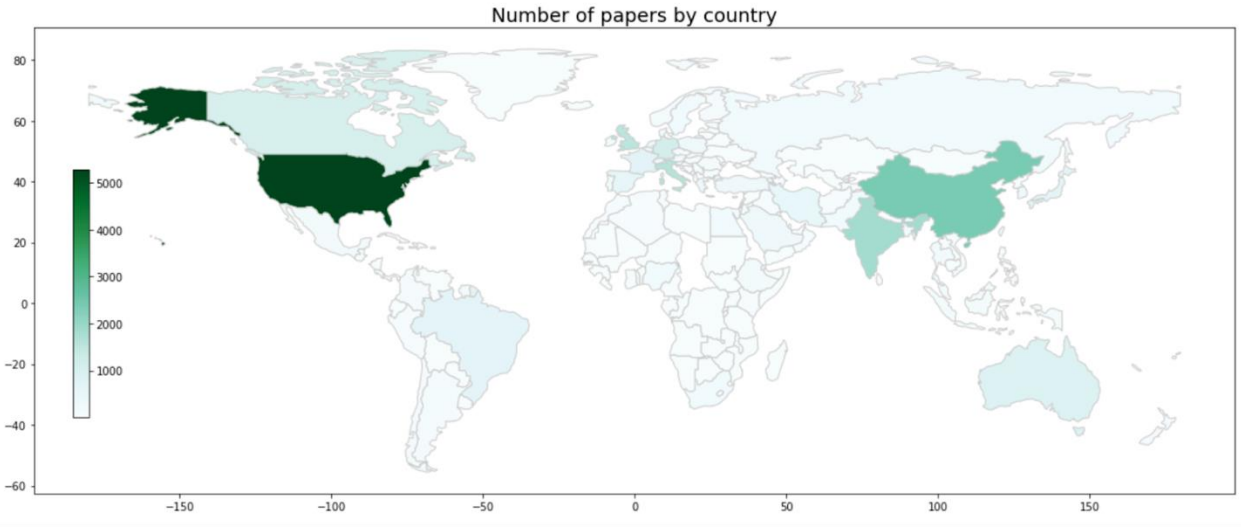


Figure 6 Number of papers by first author country of affiliation

When examining country affiliation distribution by topic (Table 1), we observe that the USA consistently maintains its position as the main generator of COVID research, with the following slots in the top 3 alternating between China and India, and western European countries (Italy, Germany and the UK) depending on topic. For all topics, the top 10 countries represent 60% or more of the total associated research. This concentration may be of particular concern for topics that are more sensitive to contextual factors and for whom lack of diversity might result in incorrect/harmful decisions. For example, vaccination uptake or methodologies for COVID studies research can be highly influenced by cultural, economic, political and other contextual factors. Given their potential to be used to inform decisions such as public policy deployment, taking this disparity in distribution into account during downstream tasks such as Question Answering is of vital importance.

Topic and cumulative proportion of paper count				
Vaccine development	Vaccination side-effects / Treatments	Vaccination efficacy (including vaccination in patients with other diseases)	Methodologies for COVID studies (e.g. statistical modelling, simulations)	Vaccine uptake and effects (by factors like age, race, etc.)
United States of America, 0.21	United States of America, 0.17	United States of America, 0.21	United States of America, 0.19	United States of America, 0.19
China, 0.36	Italy, 0.25	China, 0.3	India, 0.27	United Kingdom, 0.27

India, 0.46	China, 0.33	Germany, 0.37	China, 0.34	Italy, 0.33
Germany, 0.49	India, 0.41	Italy, 0.44	United Kingdom, 0.4	China, 0.39
United Kingdom, 0.53	Germany, 0.46	United Kingdom, 0.49	Italy, 0.46	Canada, 0.44
Italy, 0.56	United Kingdom, 0.5	France, 0.53	Australia, 0.51	India, 0.48
Canada, 0.59	Iran, 0.54	Canada, 0.56	Canada, 0.55	Germany, 0.51
Australia, 0.62	Japan, 0.58	Israel, 0.59	Germany, 0.59	Israel, 0.54
Brazil, 0.64	Canada, 0.61	Japan, 0.62	Brazil, 0.62	France, 0.56
Iran, 0.67	France, 0.63	India, 0.65	France, 0.64	Australia, 0.59

Table 1 Top 10 countries (and cumulative proportion of papers count) by topic

Further information regarding author's age, gender or ethnicity distribution was not available.

7. ANNOTATOR DEMOGRAPHIC

The training data for the sequential sentence classification task was partially hand annotated by one of the four dataset curators. The annotator is a white non-binary American graduate student at the University of Washington [20-30].

8. SPEECH SITUATION AND TEXT CHARACTERISTICS

Research papers were collected from all over the world until the date '2022-04-28'. The first level of metadata was generated at Allen institute(US) and the other data augmentation and curation was done by students at the University of Washington. The modality of the dataset is "written." Since the dataset is comprised of research papers and computer-generated values, it is categorized as scripted/edited. The topic of the dataset is academic research related to the COVID-19 vaccine. The dataset 'CORD-19-Vaccination' is collection of metadata and structured full text papers.

9. PREPROCESSING AND DATA FORMATTING

Preprocessing is divided into three phases.

- Extraction Phase : User and Context Information extraction from CORD-19
- Data Augmentation Phase: 'Language Id', 'Authors Demography', 'Keywords' and 'Topic'
- Task implementation: Question and Answering and 'Sentence Sequence Classification'

9.1 Extraction Phase: User and Context Information extraction from CORD-19

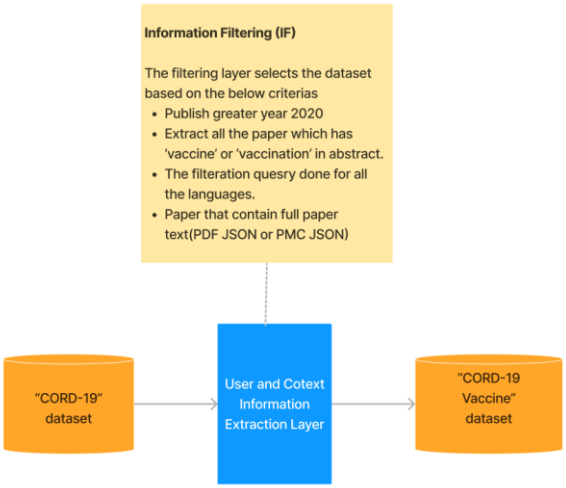


Figure 7: User and Context Information extraction

The CORD-19-Vaccination dataset is extracted from the CORD-19 dataset based on the filtering criteria below:

- Publish time: The CORD-19 metadata.csv has a column 'Publish time'. The extraction filter extracts all data where 'Publish time' is greater than or equal to '2020'.
- Pattern 'vaccine' or 'vaccination': CORD-19 metadata.csv has the columns 'Title' and 'Abstract'. All papers which had the word 'vaccine' or 'vaccination' in either the Title or the Abstract were extracted/selected. The pattern search for 'vaccine'/'vaccination' was run in all languages since there are papers in CORD-19 that are not in English. Figure 9 shows a small snippet of the code.

```
title LIKE '%مصل%' OR abstract LIKE '%مصل%' OR
title LIKE '%peyvənd%' OR abstract LIKE '%peyvənd%' OR
title LIKE '%ваксина%' OR abstract LIKE '%ваксина%' OR
title LIKE '%vacuna%' OR abstract LIKE '%vacuna%' OR
title LIKE '%vakcína%' OR abstract LIKE '%vakcína%' OR
title LIKE '%vaccine%' OR abstract LIKE '%vaccine%' OR
title LIKE '%Impfung%' OR abstract LIKE '%Impfung%' OR
title LIKE '%εμβόλιο%' OR abstract LIKE '%εμβόλιο%' OR
title LIKE '%vaccine%' OR abstract LIKE '%vaccine%' OR
title LIKE '%vacuna%' OR abstract LIKE '%vacuna%' OR
```

Figure 8: Snippet for 'vaccine/vaccination' pattern search in different languages

- Pdf_json_files / pmc_json_files : CORD-19 metdata.csv has the columns 'pdf_json_files' and 'pmc_json_files'. These columns give the path of the json files. All papers selected had the 'pdf_json_file' or 'pmc_json_files' present.
- Abstract is not null: All papers selected had the 'Abstract' column present. Our reasoning for this is that our exploratory data analysis revealed that almost all standard published papers must follow a particular template where the abstract must be present. The papers which included an abstract were mostly articles in journals. This improved the quality of our dataset as it only included standard research papers.

Since several of the 'CORD-19-Vaccination' columns are based on CORD-19 . The preprocessing for CORD-19 is explained in Wang et al. 2020 and in the Data Statement attached in the Appendix (Singh 2022).

Implementation / Code:

- The CORD-19 dataset with release date '2022-04-28', Version 110[1], was approximately 75.84 GB (Allen Institute for AI 2022). Processing data of this scale on a local computer was not possible. In order to handle data of this volume we set up a SQL / python database pipeline.
- The SQL code for the information filtering is attached in Figure 10. The code is uploaded at the GitHub repository[2]

9.2 Data Augmentation Phase: Data Augmentation Phase: 'Language Id', 'Authors Demography', 'Keywords' and 'Topic'

CORD-19-Vaccination is augmented with data that will support enhanced information retrievals and other NLP tasks. Figure 11 shows the details about implementation of each data augmented field. Classification of the 'Abstract' sentences to the labels, "Background/Objective", "Method", "Result/Conclusion". This task will be explained in more detail in section 9.2.3 Topic Modeling.

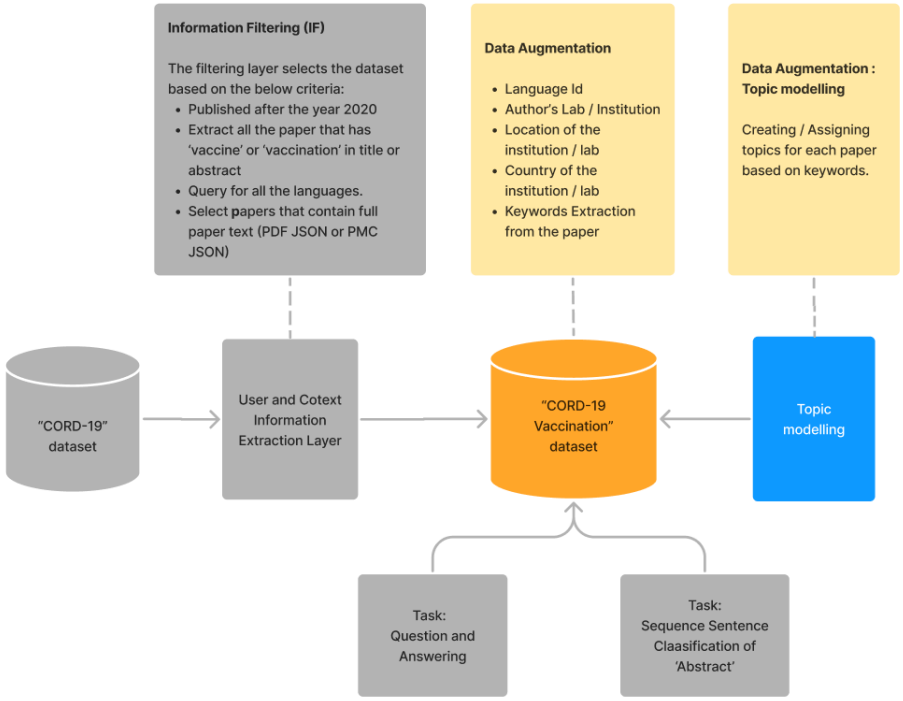


Figure 9: Data Flow - Data Augmentation

9.2.1 Language ID, Author demography, Keyword.

The CORD-19-Vaccination dataset is augmented with fields like (Language ID, Author's Lab/Institution, Location of the Institution/Lab, Country of the Institution/ Lab, and Keywords extraction from each paper.) Language ID is included in the dataset in order to support text demography. In order to establish collaboration between institutions/lab and country fields such as first Author's Lab/Institution, Location of the Institution/Lab and Country of the Institution/ Lab are included. Keyword extraction supplies the most important key words and phrases in the text. This field will support information retrieval tasks based on

[1] https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge
[2] https://github.com/manisha-Singh-UW/CORD-19-Vaccination/tree/main/data_extraction_sql

keywords. The following subsections show the details about implementation of each augmented field. Figure 12 shows the data flow for the augmented columns.

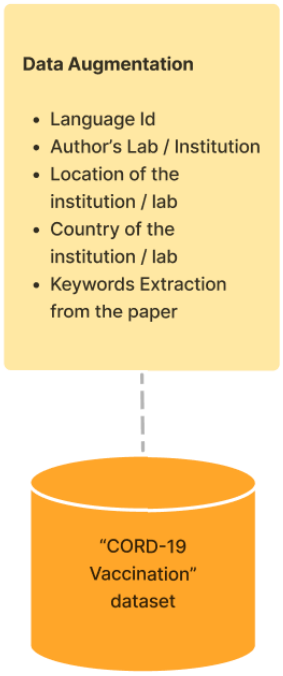


Figure 10: Dataflow for : Language ID , Author detail, Keyword augmentation

9.2.1.1 Language ID

The CORD-19 dataset did not provide any information about the text demography for each paper. Language ID will not only add this information to CORD-19-Vaccination, but will also support topic modelling by telling the model what language to use. The columns (language_id, language_id_confidence and language_id_predictions) are added to the dataset for this purpose.

Background:
For the implementation of our language identification system we used Facebook's AI fastText classifier. fastText was trained on data from Wikipedia, Tatoeba and SETimes. It is generally on par with deep learning classifiers in terms of accuracy for language identification and can recognize the following 176 languages:
['af', 'als', 'am', 'an', 'ar', 'arz', 'as', 'ast', 'av', 'az', 'azb', 'ba', 'bar', 'bcl', 'be', 'bg', 'bh', 'bn', 'bo', 'bpy', 'br', 'bs', 'bxr', 'ca', 'cbk', 'ce', 'ceb', 'ckb', 'co', 'cs', 'cv', 'cy', 'da', 'de', 'diq', 'dsb', 'dty', 'dv', 'el', 'eml', 'en', 'eo', 'es', 'et', 'eu', 'fa', 'fi', 'fr', 'frr', 'fy', 'ga', 'gd', 'gl', 'gn', 'gom', 'gu', 'gv', 'he', 'hi', 'hif', 'hr', 'hsb', 'ht', 'hu', 'hy', 'ia', 'id', 'ie', 'ilo', 'io', 'is', 'it', 'ja', 'jbo', 'jv', 'ka', 'kk', 'km', 'kn', 'ko', 'krc', 'ku', 'kv', 'kw', 'ky', 'la', 'lb', 'lez', 'li', 'lmo', 'lo', 'lrc', 'lt', 'lv', 'mai', 'mg', 'mhr', 'min', 'mk', 'ml', 'mn', 'mr', 'mrj', 'ms', 'mt', 'mwl', 'my', 'myv', 'mzn', 'nah', 'nap', 'nds', 'ne', 'new', 'nl', 'nn', 'no', 'oc', 'or', 'os', 'pa', 'pam', 'pfl', 'pl', 'pms', 'pnb', 'ps', 'pt', 'qu', 'rm', 'ro', 'ru', 'rue', 'sa', 'sah', 'sc', 'scn', 'sco', 'sd', 'sh', 'si', 'sk', 'sl', 'so', 'sq', 'sr', 'su', 'sv', 'sw', 'ta', 'te', 'tg', 'th', 'tk', 'tl', 'tr', 'tt', 'tyv', 'ug', 'uk', 'ur', 'uz', 'vec', 'vep', 'vi', 'vls', 'vo', 'wa', 'war', 'wuu', 'xal', 'xmf', 'yi', 'yo', 'yue', 'zh'] (Joulin et al. 2016).
A per figure 3: '*Language distribution in CORD-19*' gives the language distribution of 'CORD-19' dataset.
In order to capture the word 'vaccine/vaccination' in every language present in CORD-19 the query for the information extraction process was customized to search the pattern 'vaccine/vaccination' in every language as shown in section 9.1.

Implementation:
The input to the Language ID model was the text of the 'abstract' from each paper and the output was the three fields detailed below.
The Language ID = This is the language that the fastText model assigned the highest confidence score.
The Language ID confidence = This is the confidence score for the Language ID which the fastText model assigned the highest confidence score.
The Language ID prediction = This is the confidence score and Language ID for the three languages that the model assigned the highest scores.
Figure 14 includes an example of the three output fields.

lang_id	lang_id_confidence	lang_id_predictions
en	0.9167	en=0.9167, id=0.0055, fr=0.0043

Figure 11: Language ID - Sample data

The fastText model predicts the language as 'English' with a confidence level of '0.9167'. However the fastText model also gives a small confidence level to 'Indonesian' at '0.0055' and 'French' at '0.0043'. This is likely due to the medical domain including many loan words. In the example in Figure 14, the confidence level of 'English' compared to the other languages is much higher, so Language ID field is set to 'English'.

Implementation Code: The python code for language detail identification is uploaded at GitHub repository[3] .

9.2.1.2 Author's demography (lab/institution location and country)

CORD-19 dataset gave the authors name and journal in which the paper/article/journal was published. However in order to get more details regarding the author's demography, we augmented the data with authors 'lab/institution affiliation', 'lab/institution location' and 'lab/institution country'. Details on the author's demography can be used to construct a collaboration network to illustrate collaborations or coauthorship relations among institutions as in paper here.

Background:
The authors 'lab/institution affiliation', 'lab/institution location' and 'lab/institution country' was not mentioned in the CORD-19.csv metadata file. However in order to do descriptive analysis such as number of the papers contributed by each institution as done in the paper we need the institution detail related to each author of the paper. As per the paper descriptive analysis of the dataset focused on determining the research institutions of authors, geographical distribution of institutions, and collaboration among institutions from different countries and regions. The presence of the authors demography gives an idea about the paper perspective on the given research field.

Implementation:
The author detail is present in the associated json file for each paper, from which institution of affiliation, location and country were extracted. The code is attached below. The input to the code is the json file and output is the below columns corresponding to each author and paper id.

authors	*aff_lab_inst*	*aff_location*	*aff_country*
Booth, Jayaum S.; Goldberg, Eric; Barnes, Robin S.; Greenwald, Bruce D.; Sztein, Marcelo B.	*University of Maryland School of Medicine*	*postCode=21201; region=MD; settlement=Baltimore*	*USA*

Figure 12: Author detail - Sample data

[3] https://github.com/manisha-Singh-UW/CORD-19-Vaccination/tree/main/code_data_augmentation

Implementation Code: The python code for Author demography is uploaded at GitHub repository[4] .

Additionally, as the country of affiliation metadata was only available for approximately 63% of the json files, further data augmentation was carried out to extract the country of the first author via web scraping. For this process, titles of papers with missing country data were searched through Python's Google search API and the HTML source code of the webpage corresponding to the first query result was parsed using Selenium and BeautifulSoup. Scraped titles from the search query and their linked countries of affiliation were stored and subsequently validated by comparing similarity between the original CORD-19 paper title and the scraped title. Entries with a similarity below 0.4 (calculated using the Sequence Matcher module from Python's diff lib library) were excluded.

9.2.1.3 Key word from 'abstract', 'title' and 'body text'

The keywords column provides a list of key words or phrases retrieved using Yake. Yake - Yet Another Keyword Extractor. The Yake documentation is available here. Yake is an unsupervised approach for automatic keyword extraction using text features. Keywords from every text can be used further in topic modeling and also in keyword search. The list of the keywords for every paper gives an idea about the main content of the paper.

Background:
Yake is used to extract the key word from text of each paper . The reason we chose to use Yake is because it takes an unsupervised approach and is corpus, domain and language independent. Yake follows an unsupervised approach which builds upon features extracted from the text, making it applicable to documents written in different languages without the need for further knowledge. These key properties of Yake API make it perfect for our purposes.

Implementation:
Yake takes a text file as an input. This text file is a combined file of a paper's 'Title', 'Abstract' and 'body text'. The output is the list of the key words. One can customize the amount of the top key words and n-grams. For this project we took the top 20 key words and set the n-gram size as 3. Any other values were left as the default Yake values.
Figure 18 shows the result of running Yake on one of the papers in CORD-19-Vaccination. As you can see, the paper focuses on vaccinating young individuals.

keywords
young; vaccination; time; young individuals; model; vaccinating; epidemic; individuals; social; population; leisure; vaccinating young individuals; family; percent; death rate; rate; vaccine; framework; representative family; infected

Figure 13: Keyword of a paper - Sample data

Implementation Code: The python code for 'keywords' using Yake is uploaded at GitHub repository[5] .

9.2.2 Topic modelling

[4] https://github.com/manisha-Singh-UW/CORD-19-Vaccination/tree/main/code_data_augmentation

[5] https://github.com/manisha-Singh-UW/CORD-19-Vaccination/tree/main/code_data_augmentation

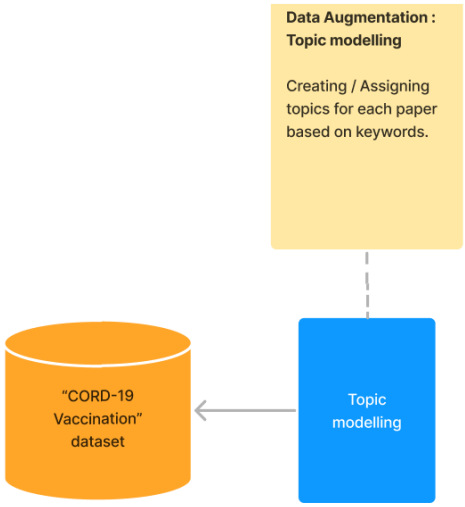


Figure 14: Data Flow for Topic modelling

Topic modelling corresponds to a series of techniques used to infer common themes within a series of documents through the generation of probabilistic models. In the context of this dataset, the Latent Dirichlet Allocation method is employed to obtain underlying topics in an unsupervised way. Generation of these topics has a twofold intention. First, it provides a comprehensive view of the dataset's recurrent paper themes and allows for easy inspection of distribution among them. Second, it serves as a quick method for users to subset the data to better fit more scoped tasks.

Background:
Latent Dirichlet Allocation (LDA) is a statistical generative model developed by Pritchard et al. which infers possible topics by generating the probabilistic distribution of a document's composition, under the assumption that each document is a mixture of different topics and the appearance of some words is more relevant for certain topics. LDA provides an efficient way to generate possible underlying topics to a set of documents, and requires no additional information apart from the documents themselves (i.e. no a priori target is required for training), which makes it ideal for exploratory tasks. It should be noted that LDA does not create a label for the topics it identifies. It is up to the user to determine the appropriate label for each topic based on manual inspection (frequently through examination of the most common words per topic) and contextual knowledge.

Implementation:
LDA is implemented through the Scikitlearn API over the collections of paper abstracts. Pre-processing is carried out to remove stop words (using the NLTK standard list for English), punctuation and words of length less than 3 are also removed. Tokenization of documents is implemented using NLTK's word tokenizer function and words are lemmatized using the same library's WordNet lemmatizer. As input to the model, a sparse matrix set to include a the word count per document of the top 3000 features is created via the Count Vectorizer implementation from Scikitlearn. Training is carried out tuning the number of topics parameter (ranging from n=5 to n=14 topics), evaluated using the Coherence score as presented by Roder et al. Through this approach, n=5 is chosen as the number of topics to train the final model.

Further evaluation of the topic labelling is done visually by performing dimensionality reduction to the topic distribution vectors of each document using t-SNE (to get 2-dimensional vectors) and plotting them colored by topic label. Figure 23 shows that topic clusters are mostly well-defined, with some small overlaps between topics such as "Vaccine uptake" and "Vaccination efficacy".

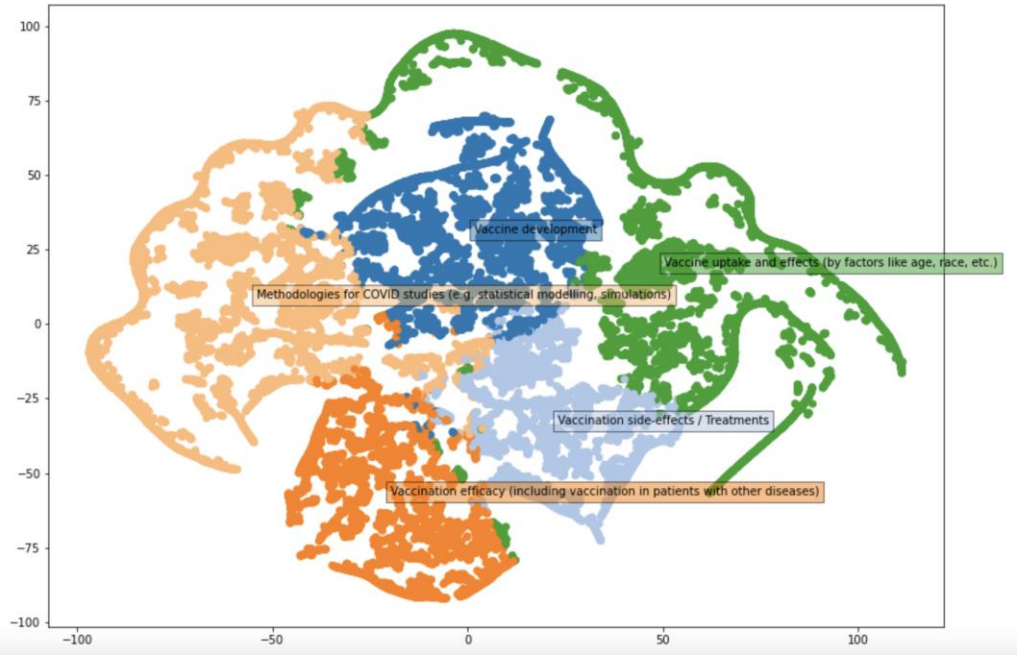


Figure 15 Clusters by topic 2-dimensional representation

Results:
Labelling of the topics is done by manual examination of the top 20 most frequent words per topic, which are also cross referenced with the title of the articles with the highest probability for the corresponding topic. Through this method, the following topics are identified within the corpus:

- Vaccine development
- Vaccination side-effects / Treatments
- Vaccination efficacy (including vaccination in patients with other diseases)
- Methodologies for COVID studies (e.g. statistical modelling, simulations)
- Vaccine uptake and effects (by factors like age, race, etc.)

9.3. Task Implementation Phase

The goal of 'CORD-19-Vaccination' dataset is to help researchers find the most relevant research papers in CORD-19 vaccine domain. The Figure 17: '*Number of journals from CORD-19 metadata readme file'* shows the trend in increase in number of journals from March 2020 till April 2022 in CORD-19 metadata.csv [6] file. At time of competition launch in March 2020, the total number of journals was 44k after two years later in April 2022, it is over 1 million rows. So, to address the information overload of CORD-19 dataset we curated 'CORD-19-Vaccination' dataset which is specific to 'vaccine' domain.

[6] https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge

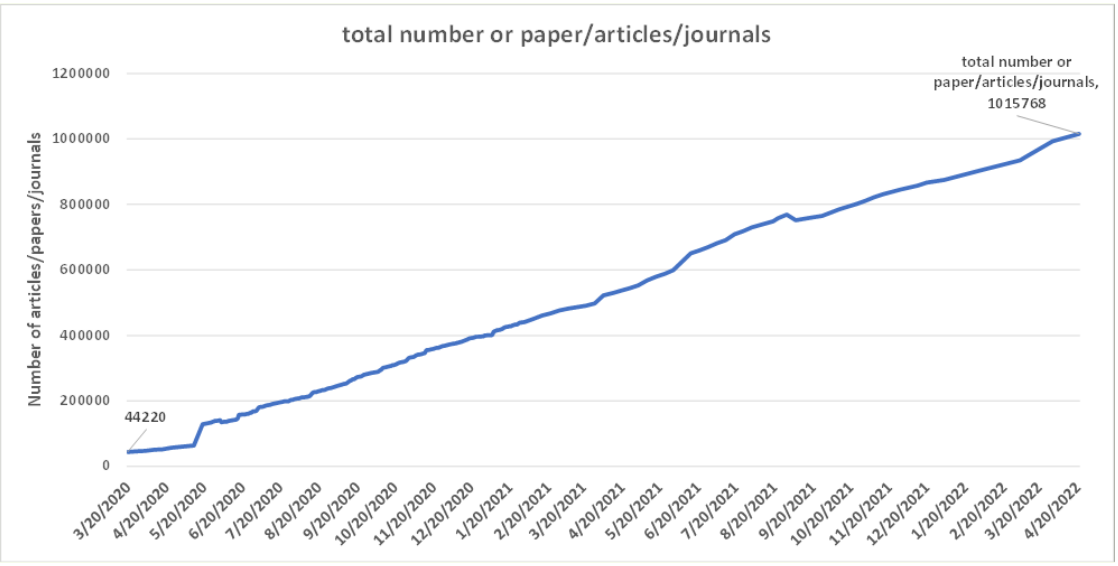


Figure 16: Number of journals from CORD-19 metadata readme file

In order to evaluate the dataset we implemented two tasks 'Question and Answering' and 'Sentence Sequence Classification'. Figure 17 – Shows the data flow for both the task and in section 9.4.1 and 9.4.2 we discuss the task implementation in detail.

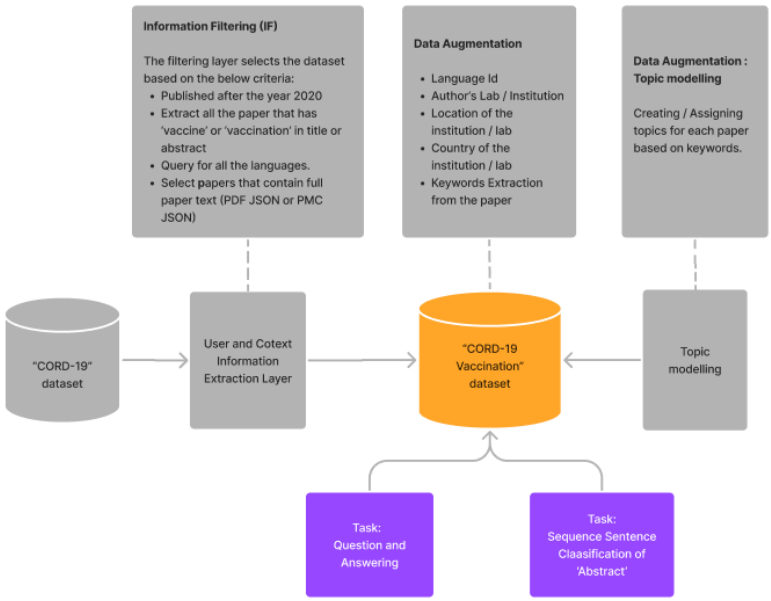


Figure 18: Data Flow: Task implementation

Abstract sentence labelling using sentence sequence classification:

The motivation of this task was from Dernoncourt and Lee's 2017 paper PubMed 200k RCT: a Dataset for Sequential Sentence Classification in Medical Abstracts. '*Automatically classifying each sentence in an abstract would help researchers read abstracts more efficiently, especially in fields where abstracts may be long, such as the medical field'* (Dernoncourt & Lee 2017).

Another paper which did a similar exercise using hand annotation on CORD-19 dataset is C*ODA-19: Using a Non-Expert Crowd to Annotate Research Aspects on 10,000+ Abstracts in the COVID-19 Open*

Research Dataset by Huang et al. "*Reliable human annotations help scientists access and integrate the rapidly accelerating coronavirus literature, and also serve as the battery of AI/NLP research, but obtaining expert annotations can be slow. We demonstrated that a non-expert crowd can be rapidly employed at scale to join the fight against COVID-19" (Huang et al. 2020).*

Taking motivation from both authors we decided to have a task in which we classify the abstracts present in the 'CORD-19-Vaccination' dataset. Our task is machine labelling using a neural network, and we validate the accuracy of the results using Dernoncourt et al.'s 2016 paper.

9.4.1 Kaggle Question and Answering task

Background:
Due to the sheer volume of data available in the CORD-19 dataset, it is impossible to run the question and answering task with the current version (111) of CORD-19. None of the top scoring solutions from the CORD-19 Kaggle challenge[7] can run using the current version of CORD-19. Our solution to this problem was to create CORD-19-Vaccination, which is an augmented subset of CORD-19 consisting of just 30k of the 1 million articles and papers in CORD-19. CORD-19-Vaccination performs well on the question and answering task.

The questions asked as a part of the Kaggle competition were all relevant for their time, but there were no questions pertaining to vaccines. As CORD-19-Vaccination focuses on vaccines, we created new vaccine-based questions for our evaluation process.

Questions:
In the absence of medical expert, we designed our questions based on a search at 'https://scholar.google.com/' related to COVID-19 vaccine as shown below:

Figure 19: COVID-19 vaccine related search at 'scholar.google.com'

Answers:
In the absence of medical expert in the field of COVID-19 vaccine, we evaluated the results of Question and Answering task based on "user-based approach" like the one outlined in Diekema et al. 2004. We based our evaluation of the quality and relevance of the papers given as answers on data on citations, views and downloads as shown in Figure 21.

Implementation:
The Question and Answering task consists of three parts: 'question', 'context', and 'answer'. The input to model is a covid-19 vaccine specific question and its context. In this implementation we are assuming that the question is contained in the context. We need to keep the context small as we have 30k journals. This

[7] https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge/discussion/148807

is done by selecting the papers similar to the answers, using 'Okapi BM25' (Wikipedia, 2022). Okapi BM25 is a ranking function used by search engines to estimate the relevance of the document for a given search query. For each question and context, we are using "Huggingface transformer library"[8] to predict the answer (Wolf et al., 2020). We have used the pretrained model 'bert-large-uncased-whole-word-masking finetuned-squad.' The solution for this task was customized for 'CORD-19-vaccination' dataset which is inspired by Besomi 2020's Kaggle notebook.

Output:

Figure 20 is the output of the 'Question and Answering' task

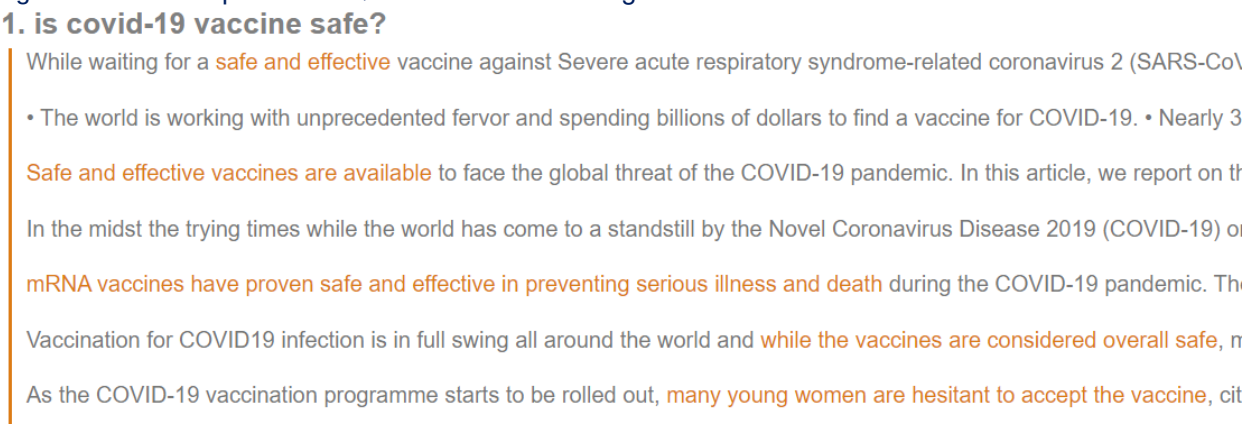


Figure 20: Question and Answering output

Figure 27 gives a list of papers as answers to the question 'Is Covid-19 vaccine safe? '. According to the "user-based approach" of evaluation we can say that the papers in the result seem relevant, as most of the papers were recently published.

Question	Papers	Citation	Viewed	Downloads
Is Covid-19 vaccine safe?	10.1038/s41577-021-00525-y	111		
	10.1016/j.puhe.2020.05.007	9		
	10.1093/jlb/lsaa024	3	1146	435
	10.3390/vaccines10020298		627	
	10.1111/jdv.17499	3		
	10.1111/dth.15146	6		

Figure 21: Answers 'user-based' analysis

Evaluation of 'CORD-19-Vaccination' dataset:

Why 'CORD-19-Vaccination' dataset is better that 'CORD-19' for vaccination related 'Question and Answering' task:

1. As written in the paper Mollá and Vicedo 2007 "*QA systems are especially useful in situations in which a user needs to know a very specific piece of information and does not have the time—or*

[8] https://github.com/huggingface/transformers

just does not want—to read all the available documentation related to the search topic in order to solve the problem at hand". The 'CORD-19-Vaccination' dataset is more domain specific and therefore the results are more accurate and specific to the task. To make dataset more easily available, we curated the dataset by filtering on the first level if the 'abstract' or 'title' contains 'vaccine' or 'vaccination.' This makes the response time of the application faster for any question, since now the context search is done on 30k rows rather than on 1 million unfiltered rows of CORD-19.

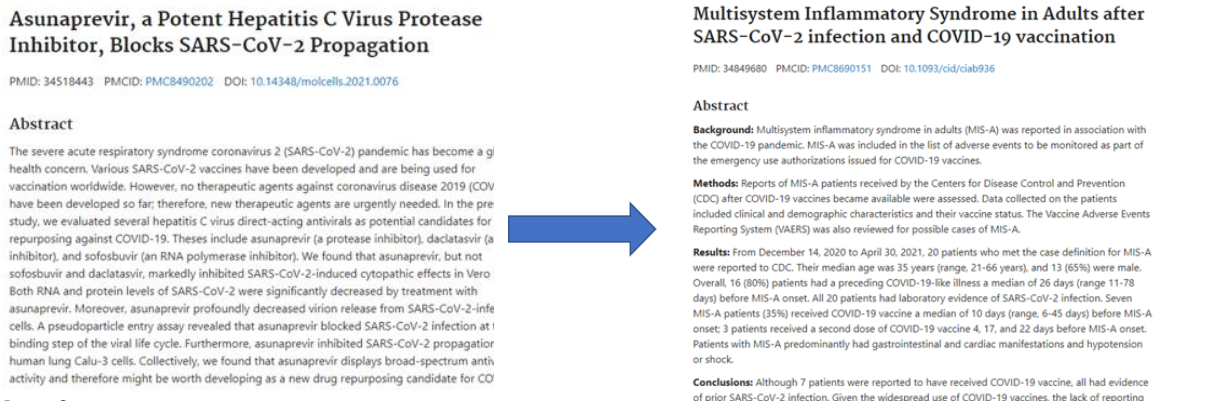
2. 'CORD-19-Vaccination' was augmented with fields 'Language ID', 'Authors affiliation based on Lab/institution, location, country' and 'Topic', 'Keywords' and 'Abstract labelled sentences'. These fields are not present in 'CORD-19'. Researchers using 'CORD-19-Vaccination' can make use of these augmented fields for better answers.
3. The column 'keyword' in 'CORD-19-Vaccination' dataset is the list of keywords extracted from the body of the text papers using 'Yake', so if we extract the context using 'title' + 'abstract' + 'Keyword', the answer should be more accurate.
4. As per the paper Diekema 2004, the 'CORD-19-Vaccination' dataset has the maximum coverage related to 'vaccine' and 'up to datedness'. Citing from the paper Diekema 2004, "*the documents in the e-Query database were useful, but Google is much faste*r". The curated 'CORD-19-Vaccination' database gives all the relevant papers nearly as fast as and more specifically than Google.

9.4.2 Sequential Sentence Classification Task

Background:

Text classification is a very important task in Natural Language Processing (NLP) where a label or class is assigned to a text. There are many applications of this task such as Sentiment Analysis, Question Answering, Spam filtering, Sorting emails by topics, and Genre Classification.

In the current task, the focus is on the classification of sentences in medical abstracts. The sentences in the abstracts appear in a sequence therefore this task is called "Sequential Sentence Classification Task", to distinguish it from general text or sentence classification that does not consider the context (for example, previous or next sentences, or the position of the sentence in a block of text).



Asunaprevir, a Potent Hepatitis C Virus Protease Inhibitor, Blocks SARS-CoV-2 Propagation

PMID: 34518443 PMCID: PMC8490202 DOI: 10.14348/molcells.2021.0076

Abstract

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) pandemic has become a gl health concern. Various SARS-CoV-2 vaccines have been developed and are being used for vaccination worldwide. However, no therapeutic agents against coronavirus disease 2019 (COV have been developed so far; therefore, new therapeutic agents are urgently needed. In the pre study, we evaluated several hepatitis C virus direct-acting antivirals as potential candidates for repurposing against COVID-19. Theses include asunaprevir (a protease inhibitor), daclatasvir (a inhibitor), and sofosbuvir (an RNA polymerase inhibitor). We found that asunaprevir, but not sofosbuvir and daclatasvir, markedly inhibited SARS-CoV-2-induced cytopathic effects in Vero Both RNA and protein levels of SARS-CoV-2 were significantly decreased by treatment with asunaprevir. Moreover, asunaprevir profoundly decreased virion release from SARS-CoV-2-infe cells. A pseudoparticle entry assay revealed that asunaprevir blocked SARS-CoV-2 infection at t binding step of the viral life cycle. Furthermore, asunaprevir inhibited SARS-CoV-2 propagatior human lung Calu-3 cells. Collectively, we found that asunaprevir displays broad-spectrum anti activity and therefore might be worth developing as a new drug repurposing candidate for CO

Paper Source:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8490202/

Multisystem Inflammatory Syndrome in Adults after SARS-CoV-2 infection and COVID-19 vaccination

PMID: 34849680 PMCID: PMC8690151 DOI: 10.1093/cid/ciab936

Abstract

Background: Multisystem inflammatory syndrome in adults (MIS-A) was reported in association with the COVID-19 pandemic. MIS-A was included in the list of adverse events to be monitored as part of the emergency use authorizations issued for COVID-19 vaccines.

Methods: Reports of MIS-A patients received by the Centers for Disease Control and Prevention (CDC) after COVID-19 vaccines became available were assessed. Data collected on the patients included clinical and demographic characteristics and their vaccine status. The Vaccine Adverse Events Reporting System (VAERS) was also reviewed for possible cases of MIS-A.

Results: From December 14, 2020 to April 30, 2021, 20 patients who met the case definition for MIS-A were reported to CDC. Their median age was 35 years (range, 21-66 years), and 13 (65%) were male. Overall, 16 (80%) patients had a preceding COVID-19-like illness a median of 26 days (range 11-78 days) before MIS-A onset. All 20 patients had laboratory evidence of SARS-CoV-2 infection. Seven MIS-A patients (35%) received COVID-19 vaccine a median of 10 days (range, 6-45 days) before MIS-A onset; 3 patients received a second dose of COVID-19 vaccine 4, 17, and 22 days before MIS-A onset. Patients with MIS-A predominantly had gastrointestinal and cardiac manifestations and hypotension or shock.

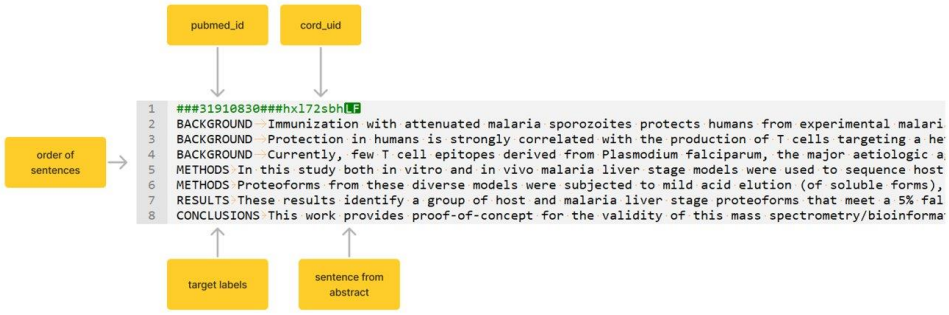
Conclusions: Although 7 patients were reported to have received COVID-19 vaccine, all had evidence of prior SARS-CoV-2 infection. Given the widespread use of COVID-19 vaccines, the lack of reporting of MIS-A associated with vaccination alone, without evidence of underlying SARS-CoV-2 infection, is reassuring.

Paper Source:
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8690151/

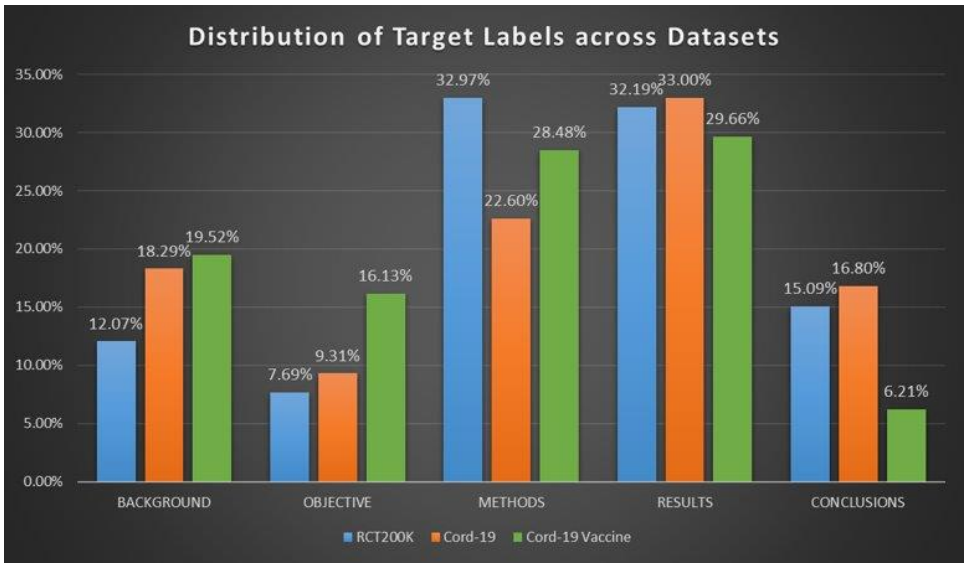
Unstructured block-of-text abstracts makes it difficult to quickly access the information of interest	- Structured abstracts (text organized into semantic headings such as Background, Methods, Results, and Conclusions) makes it easy to quickly locate relevant information. - This structure facilitates multiple downstream tasks such as automatic text summarization, information extraction, and information retrieval. - This task is based on the paper "PubMed 200k RCT: a Dataset for Sequential Sentence Classification in Medical Abstracts"

Implementation:

Training Data Sample: The data for training the model for this task is obtained from the CORD-19 dataset itself. 11.58% of the abstracts from the CORD-19 dataset (approximately 117k samples) were found to have abstracts structured with semantic headings. Similarly, 14.66% of the records in the CORD-19-Vaccine dataset (4294 samples) were found to have abstracts structured with semantic headings. These records were split into test and validation datasets for model training. A single data sample contains information on target labels, sentence from abstract, and order of sentences as shown below, and compatible with "PubMed 200k RCT: a Dataset for Sequential Sentence Classification in Medical Abstracts". The pubmed_id, cord_uid fields are available as comments and are not inputs to the training model. As per the guidance of the PubMed 200k RCT paper, numbers from the dataset have been replaced with the @ sign.

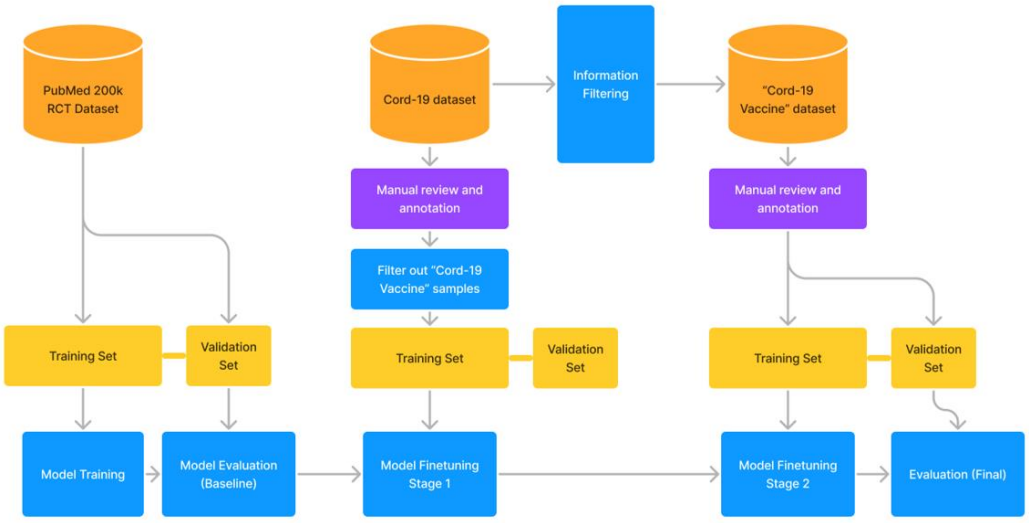


The distribution of various target labels are shown across datasets. It is important to note that the percentage of OBJECTIVE labels is quite high (at 16.13%) in the CORD-19-Vaccination dataset, while the percentage of CONCLUSION labels is quite low (at 6.21%) compared to PubMed RCT200k and CORD-19 datasets. These distributions are likely to impact the model predictions since the model is trained on all three datasets.

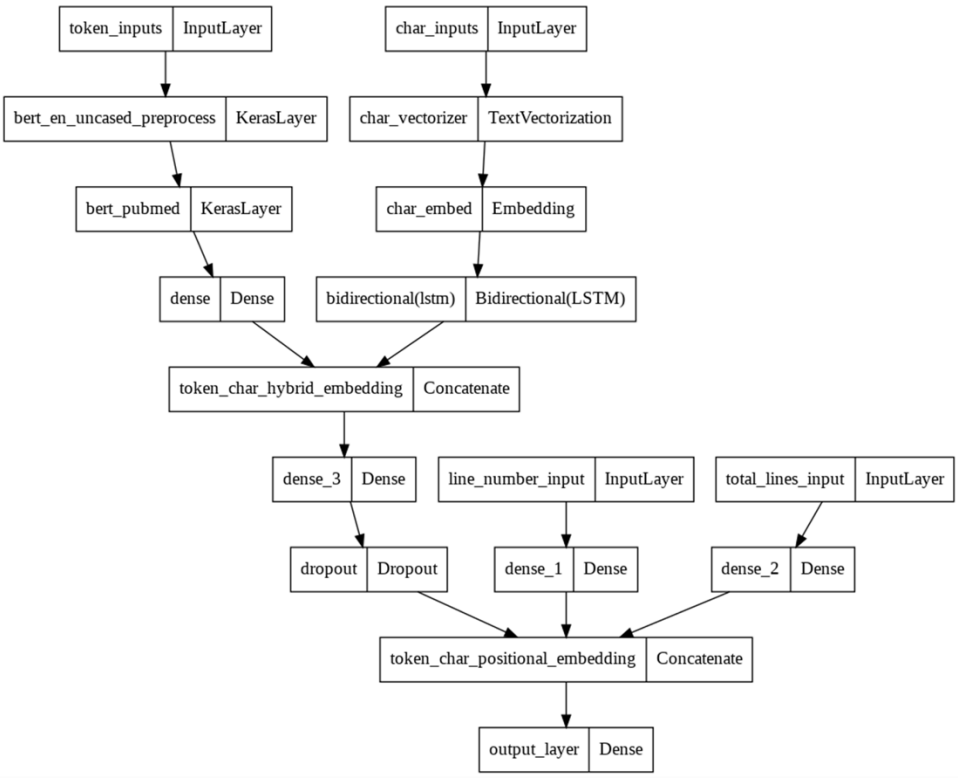


Task Workflow:

The workflow below shows the sequence of tasks performed during the training and subsequent fine-tuning of the model. This particular workflow was chosen to allow coarse-grained to fine-grained model training.



The model architecture used for training is based on the https://doi.org/10.48550/arXiv.1612.05251 paper. A pre-trained and frozen BERT-PubMed layer has been used to improve performance. The original training/validation data split of PubMed 200k RCT dataset was used fir the initial round of training. For fine tuning, a random split of 70-30 was used for stage 1 and split of 50-50 was used for stage 2. The model training was performed initially with a learning rate of 1e-4, which was reduced to 1e-4 for fine tuning. A system based on Nvidia Tesla P100 GPU was used for training.



Output:

Here are the performance metrics of this model on the CORD-19-Vaccine dataset

Accuracy	F1-score	Precision	Recall
0.7617	0.75691	0.75687	0.76179

Evaluation:

Most-wrong predictions: Additional evaluation was performed by manually reviewing the most wrong predictions. Some patterns that we found in these predictions are short sentences consisting of just a few words were incorrectly predicted, and ungrammatical or ambiguous sentences were misclassified.

target	text	Line number	Total lines	prediction	Prediction probability
BACKGROUND	we investigated if people's response to the official recommendations during the covid-@ pandemic is associated with conspiracy beliefs related to covid-@, a distrust in the sources providing information on covid-@, and an endorsement of complementary and alternative medicine (cam).	0	11	OBJECTIVE	0.993720651
RESULTS	@% (n = @).	10	22	METHODS	0.992311656
METHODS	we applied natural language processing (nlp) and thematic analysis to understand public opinions,	4	15	OBJECTIVE	0.991290927

	experiences, and issues with respect to the covid-@ pandemic using social media data.				
RESULTS	systemic adverse events were more common after the second vaccination, particularly with the highest dose, and three participants (@%) in the @-µg dose group reported one or more severe adverse events.	8	11	METHODS	0.989628315
BACKGROUND	describe nursing faculty and student nurse factors associated with covid- @ vaccine readiness.	2	8	OBJECTIVE	0.983674884
RESULTS	neutralizing antibodies were detectable within seven to @ days following disease onset, with levels increasing until days @–@ before levelling and then decreasing, but titres were lower in those with asymptomatic or clinically mild disease.	13	22	METHODS	0.982594848
BACKGROUND	&aims: liver transplant (lt) recipients or other immunocompromised patients were not included in the registration trials of vaccine studies against sars-cov-@	0	15	OBJECTIVE	0.980129063
OBJECTIVE	design.	6	26	METHODS	0.979951441
BACKGROUND	we aim to present a comprehensive research protocol that will generate epidemiological, sociological and anthropological data about the covid-@ epidemic in burkina faso, a landlocked country in west africa with scarce resources.	2	8	OBJECTIVE	0.979436338
METHODS	patient demographics and clinical information including presenting symptoms, time of symptom onset, time of diagnosis and laterality.	3	12	OBJECTIVE	0.978573859

Figure 22: Sequential sentence classification incorrect predictions

The following is the Confusion Matrix plotted using scikit learn. The matrix on the left shows the raw numbers of label distribution, while the matrix on the right is normalized on the “true” labels.

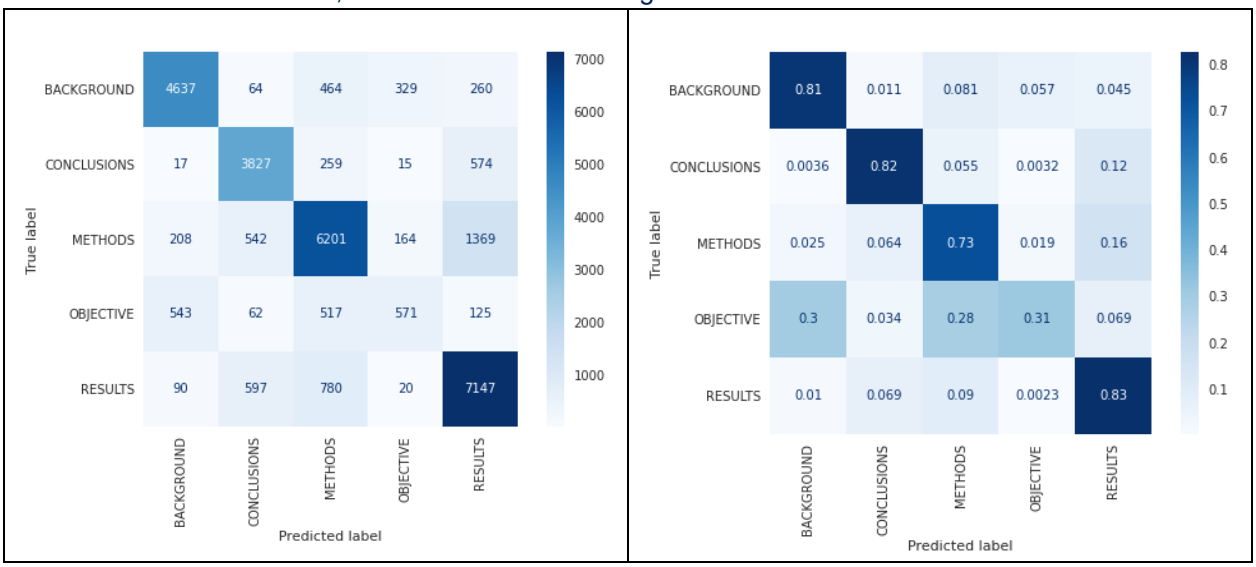


Figure 23: Confusion matrices for sequential sentence classification

We can observe from the confusion matrix that the OBJECTIVE label was often confused with BACKGROUND and METHODS. Similarly, METHODS label was often confused with RESULTS.

10. CAPTURE QUALITY

Below are the sections stating the data quality issues and the reasons behind the issues. Some were due to design selection, while some were due to missing data in the source dataset.

- For all the columns derivative from the source dataset that is ‘CORD-19’ please refer to Data statement of ‘CORD-19’ attached in the Appendix.

- The column lang_id mentions ‘en’ even if the value of the column lang_id_predictions are (en=0.2938, fr=0.1759, es=0.0769). In order to address these data quality issues, we need to deep dive into the ‘abstract’ or the ‘body text’.

- As described in section 9.2.1.2, the original dataset presented a significant completion gap (~37%) for the author’s country of affiliation. Additionally, country data showed inconsistencies both in terms of abbreviation (e.g. United States of America / USA) and spelling (e.g. USA / U.S.A). These problems were partially addressed through a web scraping process to improve completion, followed by a manual inspection phase and automated similarity matching against an standardized set of country names.

- As mentioned in section 9.4.2 and shown in figure 23, the sequential sentence classification is not 100% accurate. Most of the labels are accurate at least 80% of the time, but the Objective label is only correct roughly 31% of the time. These confusion matrices were created from running the code on only the hand annotated data, so are not accurate for the dataset as a whole, but it can be extrapolated from this data that the labels on the final dataset are not completely accurate.

11. LIMITATIONS

The dataset is specific to a particular domain. The curation rationale of the dataset provides details of the application scenarios. The CORD-19 dataset, on which the CORD-19-Vaccination is derived, is based on published academic research, therefore the characteristics of the text in the data is highly scientific, with a focus on the medical domain, and uses a formal written down structure. These characteristics must be considered when designing tasks using this dataset. For instance, the language used in the dataset may not be suitable for models used for a conversational style of text, or to compare against transcribed text from conversational speech.

12. METADATA

Below detail in the section gives the ‘License’ and citation detail

12.1 License:

As mentioned in the exercise requirement: (*A license (and appropriate attribution) for the new dataset you create should be consistent with the datasets you draw from*.). ‘CORD-19-Vaccination’ dataset is extracted from ‘CORD-19’ dataset, so ‘CORD-19-Vaccination’ dataset also follows all the license that is followed by ‘CORD-19’
Below are the license details for the CORD-19 dataset.

Dataset license link : https://ai2-semanticscholar-cord-19.s3-us-west-2.amazonaws.com/2020-03-13/COVID.DATA.LIC.AGMT.pdf

Open access license: The PMC Public Health Emergency Covid-19 Initiative expanded access to Covid-19 literature by publishers to make coronavirus-related papers discoverable and accessible through PMC under open access license terms that allow for reuse and secondary analysis.

Covid-19 open access licenses: Publishers, such as Elsevier and Springer Nature to provide full text coverage of relevant papers available in their back catalog; these papers are made available under special Covid-19 open access licenses.

Open licenses include :

- Creative Commons (CC),
- publisher-specific COVID-19 licenses
- identified as open access through DOI lookup in the Unpaywall database

12.2 How to Cite:

Since CORD-19-Vaccination dataset is extracted from 'CORD-19', when referring to the dataset in general, cite the paper (Wang et al. 2020). For 'CORD-19-Vaccination' cite GitHub Repository[9]

Singh M., Sharma D., Ma A., Tyree B. (June 2022). CORD-19-Vaccination: The COVID-19 Open Research Dataset Vaccination Subset. Unpublished manuscript. University of Washington (UW)

12.3 Errata:

For any question or concerns regarding 'CORD-19-Vaccination' , please mail to Divy Sharma(divy@uw.edu), Manisha Singh(manishas@uw.edu), Alonso Ma(amatake@uw.edu) and Bridget Tyree (btyree@uw.edu)

13. DISCLOSURES AND ETHICAL REVIEW

The dataset curation and data statement is part of course work 575 A – 'Value sensitive data processing' exercise 3.

14. OTHER

Below are some sections that in my view are useful in reference to CORD-19-Vaccination datasets.

14.1 Kaggle challenge

The Kaggle site, has a challenge based on the CORD-19. Any challenge that is relevant to CORD-19 should also be relevant to 'CORD-19-Vaccination' dataset. Obviously, the result will not be the same. The reason I am mentioning because the task attached in target table is a great way to look at the dataset and tasks that can be done using CORD-19-Vaccination dataset.

14.2 Reference

Further reference on CORD-19 (Wang et al. 2020). CORD-19 data statement refer to Appendix here.

[9] https://github.com/manisha-Singh-UW/CORD-19-Vaccination

14.3 GitHub Repository

The code and the dataset is uploaded at repo - https://github.com/manisha-Singh-UW/CORD-19-Vaccination

15. GLOSSARY

- COVID-19 : Coronavirus disease 2019 (COVID-19) is the official name given by the World Health Organization (WHO) to the disease caused by SARS-CoV-2, the new coronavirus that surfaced in Wuhan, China in 2019 and spread around the globe.
- Yake : Yet another key word extractor. Documentation here
- LDA : ()
-

16. ABOUT THIS DOCUMENT

A data statement is a characterization of a dataset that provides context to allow developers and users to better understand how experimental results might generalize, how software might be appropriately deployed, and what biases might be reflected in systems built on the software.

This data statement was written based on the template for the Data Statements Version 2 Schema. The template was prepared by Angelina McMillan-Major, Emily M. Bender, and Batya Friedman and can be found at http://techpolicylab.uw.edu/data-statements.

17. REFERENCES

(*COVID DATASET LICENSE, 2019).* COVID DATASET LICENSE AGREEMENT. Retrieved from https://ai2-semanticscholar-cord-19.s3-us-west-2.amazonaws.com/2020-03-13/COVID.DATA.LIC.AGMT.pdf

(*GROBID,* 2008--2022). GROBID. Online: GitHub. Retrieved from https://github.com/kermitt2/grobid

(*You have a system*, 2022). *You have a system using CORD-19? Let us know!*. Retrieved from https://docs.google.com/forms/d/e/1FAIpQLSdGhG8A_f1sbDQcm2m_OhCrRfjVkpx99b4oerPCFBuj_dV5AA/viewform

Allen Institute for AI, Goldbloom, A., Lin, P., Mooney, P., Schoenick, C., Kohlmeier, S, Devrishi, Bozsolik, T., Hammer, B. (2022). COVID-19 Open Research Dataset Challenge (CORD-19). Retrieved from https://www.kaggle.com/datasets/allen-institute-for-ai/CORD-19-research-challenge?datasetId=551982&sortBy=dateCreated

Besomi, J. (2020). *A qa model to answer them all.* Retrieved from https://www.kaggle.com/code/jonathanbesomi/a-qa-model-to-answer-them-all/comments

Campos, R., Mangaravite, V., Pasquali, A., Jatowt, A., Jorge, A., Nunes, C. and Jatowt, A. (2020). YAKE! Keyword Extraction from Single Documents using Multiple Local Features. In Information Sciences Journal. Elsevier, Vol 509, pp 257-289. pdf

Campos R., Mangaravite V., Pasquali A., Jorge A.M., Nunes C., and Jatowt A. (2018). A Text Feature Based Automatic Keyword Extraction Method for Single Documents. In: Pasi G., Piwowarski B.,

Azzopardi L., Hanbury A. (eds). Advances in Information Retrieval. ECIR 2018 (Grenoble, France. March 26 – 29). Lecture Notes in Computer Science, vol 10772, pp. 684 - 691. pdf

Campos R., Mangaravite V., Pasquali A., Jorge A.M., Nunes C., and Jatowt A. (2018). YAKE! Collection-independent Automatic Keyword Extractor. In: Pasi G., Piwowarski B., Azzopardi L., Hanbury A. (eds). Advances in Information Retrieval. ECIR 2018 (Grenoble, France. March 26 – 29). Lecture Notes in Computer Science, vol 10772, pp. 806 - 810. pdf

Cohan, A., Feldman, S., Beltagy, I., Downey, D., Weld, D.S. (2020). SPECTER: Document-level Representation Learning using Citation-informed Transformers. 2004.07180. Retrieved from the arXiv database.

Creative Commons. (2022). *Creative Commons*. Retrieved from https://creativecommons.org/

Dernoncourt, F., Lee, J.Y., Szolovits, P. (2016). Neural Networks for Joint Sentence Classification in Medical Paper Abstracts. 1612.05251. Retrieved from the arXiv database.

Dernoncourt, F., & Lee, J. Y. (2017). Pubmed 200k rct: a dataset for sequential sentence classification in medical abstracts. arXiv. Retrieved from https://arxiv.org/abs/1710.06071 doi: 10.48550/ARXIV.1710.06071

Diekema, A., Yilmazel, O., & Liddy, E. (2004, 01). Evaluation of restricted domain question-answering systems. Center for Natural Language Processing.

Elsevier. (2021). *Novel Coronavirus Information Center*. Retrieved from https://www.elsevier.com/connect/coronavirus-information-center

Huang, T.-H. K., Huang, C.-Y., Ding, C.-K. C., Hsu, Y.-C., & Giles, C. L. (2020). Coda-19: Using a non-expert crowd to annotate research aspects on 10,000+ abstracts in the covid-19 open research dataset. arXiv. Retrieved from https://arxiv.org/abs/2005.02367 doi: 10.48550/ARXIV.2005.02367

IBM. (2021). *IBM Watson Discovery*. Retrieved from https://www.ibm.com/cloud/watson-discovery

Joulin, A., Grave, E., Bojanowski, P., Douze, M., J´egou, H., & Mikolov, T. (2016). FastText.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*

Lo, K., Wang, L.L., Neumann, M., Kinney, R., Weld, D. (July 2020). The Semantic Scholar Open Research Corpus. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics. Retrieved from https://github.com/allenai/s2orc-doc2json

Moll´a, D., & Vicedo, J. L. (2007, 03). Question Answering in Restricted Domains: An Overview. Computational Linguistics, 33 (1), 41-61. Retrieved from https://doi.org/10.1162/coli.2007.33.1.41 doi: 10.1162/coli.2007.33.1.41

National Center for Biotechnology Information. (2021). *Journal Article Tag Suite*. Retrieved from https://jats.nlm.nih.gov/index.html

National Library of Medicine. (2021). *Public Health Emergency COVID-19 Initiative*. Retrieved from https://www.ncbi.nlm.nih.gov/pmc/about/covid-19/

Open Review. (2020). *CORD-19: The COVID-19 Open Research Dataset*. Retrieved from https://openreview.net/forum?id=0gLzHrE_t3z

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). *Inference of Population Structure Using Multilocus Genotype Data. Genetics, 945-959.*

Röder, M., Both, A., & Hinneburg, A. (2015). *Exploring the Space of Topic Coherence Measures. New York: Association for Computing Machinery.*

Semantic Scholar. (2022). *Semantic Scholar Academic Graph API*. Retrieved from https://www.semanticscholar.org/product/api (Pritchard, Stephens, & Donnelly, 2000)

Singh M., (2022). CORD-19: Data Statement. Unpublished manuscript. University of Washington (UW)

Springer Nature. (2021). *Coronavirus (COVID-19) Research Highlights*. Retrieved from https://www.springernature.com/gp/researchers/campaigns/coronavirus

Unpaywall. (2018). *Unpaywall*. Retrieved from https://unpaywall.org/

Wang, L.L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Burdick, D., Eide, D., Funk, K., Katsis, Y., Kinney, R.M., Li, Y., Liu, Z., Merrill, W., Mooney, P., Murdick, D.A., Rishi, D., Sheehan, J., Shen,

Z., Stilson, B., Wade, A.D., Wang, K., Wang, N.X.R., Wilhelm, C., Xie, B., Raymond, D.M., Weld, D.S., Etzioni, O. & Kohlmeier, S. (July 2020). CORD-19: The COVID-19 Open Research Dataset. *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*. Online: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/2020.nlpcovid19-acl.1

Wikipedia. (2022). *Okapi bm25*. Retrieved from https://en.wikipedia.org/wiki/OkapiB M 25

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., . . . Rush, A. M. (2020, October). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical*

World Health Organization. (2022). Global research on coronavirus disease (COVID-10). Retrieved from https://www.who.int/emergencies/diseases/novel-coronavirus-2019/global-research-on-novel-coronavirus-2019-ncov

18. APPENDIX

CORD-19 Data statement: CORD-19 data statement is uploaded

here: https://github.com/manisha-Singh-UW/CORD-19-Vaccination/tree/main/Datastatement