**Dataset Report: Overview and Analysis**

**Full Name: Akhona Stafane**

**23August 2024**

**Dataset Overview**

**Dataset Name:** The COVID-19 Open Research Dataset (CORD-19)

**Size and Format:**

* **Size of Last Published Version:** size: 18.7Gb, md5: c557069e, sha1: dd2c32bc
* **Format:** S2ORC JSON format
* **Number of Records:** 140K papers
* **Number of Features:** Numerous (e.g., Text, Metadata, Countries, etc.)

**Age:**

* **Date:** First published version (2020-03-13) - Last published version (2022-06-02)
* **Collection Frequency:** Daily

**Brief Overview:**

The CORD-19 dataset includes publications and preprints related to COVID-19, SARS, and MERS. It was first released on March 16, 2020, by the Allen Institute for AI (AI2) in collaboration with several partners, including The White House Office of Science and Technology Policy, the National Library of Medicine, the Chan Zuckerberg Initiative, Microsoft Research, Kaggle, and Georgetown University’s Center for Security and Emerging Technology. Initially, the dataset contained 28,000 papers and quickly expanded to 140,000 papers in the following weeks.

The dataset is built from papers and preprints gathered from various archives and processed through the Semantic Scholar search engine. Metadata is standardized, and duplicates are removed to ensure consistency. Over 50% of the papers in CORD-19 have their full text extracted through a processing pipeline. The dataset is updated daily to include new publications and maintain its relevance.

Since the release of CORD-19 on May 12, 2020, selected HTML table parses have been added to the dataset. These tables contain important numerical and descriptive data, such as sample sizes and results. A specialized PDF table processing pipeline extracts and understands tables from documents using IBM Watson Discovery’s Smart Document Understanding (SDU). This technology converts PDFs into text-based HTML representations, identifying structures like tables and formatting information. The tables are then annotated with semantic details, such as headers and captions.

A total of 188,000 tables from 54,000 documents were extracted. Of these, 33,000 tables were successfully matched to 19,000 full-text documents in CORD-19. Matches are identified based on caption similarity, with a Jaccard similarity score of 0.9 or higher. Preliminary analysis indicates that mismatches often occur due to differences in table captions between parsing schemes.

**Collection and Processing Procedures:**

* **Collection:** The data is sourced from Papers and preprints from several archives. Data collection involves integrating papers and preprints from several sources such as Word Health Organisation, PubMed Central, PubLMed, and others. These papers and preprints are ingesting through the Semantic Scholar literature search engine. A paper is defined as the base unit of published knowledge, and a preprint as an unpublished but publicly available counterpart of a paper.
* **Processing:** The raw data undergoes cleaning to remove errors and inconsistencies. This includes handling duplicates and standardising metadata. paper documents are processed through the pipeline established in Lo et al.(2020) to extract full text.
* Specifically, The key steps followed are:

**Ingestion:**

Sources: Papers in CORD-19 are sourced from PubMed Central (PMC), PubMed, the World Health Organization’s Covid-19 Database, and preprint servers bioRxiv, medRxiv, and arXiv, and are ingested through the Semantic Scholar literature search engine. Papers have metadata and documents with the actual content. To retrieve the papers a query that searches using the below words is used:

"COVID" OR "COVID-19" OR "Coronavirus" OR "Corona virus" OR "2019-nCoV" OR "SARS-CoV" OR "MERS-CoV" OR "Severe Acute Respiratory Syndrome" OR "Middle East Respiratory Syndrome"

Metadata: Each paper includes bibliographic details like title, authors, and publication venue, along with unique identifiers such as DOI, PubMed IDs, WHO Covidence #, and MAG identifier

Documents: Papers may come with associated documents, including PDFs, XMLs, or physical print-outs.

**Processing:**

Harmonization and Deduplication: We standardize and remove duplicate metadata to create a unified and clean dataset. Matadata is harmonised and depluplicated according to the following steps:

* 1. Cluster papers using paper identifiers
  2. Select canonical metadata for each cluster
  3. Filter clusters to remove unwanted entries

Full Text Parsing: For papers with accessible documents and open access licenses, structured parses of the full text are generated. Most papers are associated with one or more PDFs.To extract full text and bibliographies from each PDF, the PDF parsing pipeline created for the S2ORC dataset is used.

**Output:**

Harmonized Metadata: Cleaned and standardized metadata for easy integration and analysis.

Structured Full Text: Parsed text from documents where available, allowing for detailed content analysis.

Numerical Data: Numerical Data from the parsed and matched tables consisting of categorical data, descriptive statistics.

**Analysis According to the V’s of Big Data**

1. **Volume:**

The dataset’s volume is substantial at 18.7 GB, encompassing 140K papers with over 72K full texts and preprints on COVID-19 and related historical coronaviruses such as SARS and MERS. This large volume enables connecting the machine learning community with biomedical domain experts and policy makers in the race to identify effective treatments and management policies for COVID-19.

1. **Variety:**

The dataset is rich with variety of data such as: Harmonized Metadata Structured Full Text:

* + **Text Data:** Metadata and Full text
  + **numeric and descriptive information:** sample sizes and results, number of deaths, number of cases
  + **Categorical Data:** Gender
  + **Geospatial Data:** Countries where research is conducted.

This variety allows for multi-faceted analysis, such as examining the distribution of papers per year. Analysis can also show the countries where most of the studies are conducted.

1. **Veracity:**

The veracity of the dataset is high due to the reputable sources of data (World Health Organisation). The papers are also published which makes them reliable sources regarding Covid-19. However, data validation procedures are in place to ensure accuracy and address potential inconsistencies or errors. Common issues which include duplications in metadata were handled.

1. **Velocity:**

Hundreds of new COVID-19 publications are released daily. To remain relevant, the CORD-19 dataset is updated every day. The continuous influx of new research necessitates frequent updates to ensure the dataset reflects the most current information. The processing pipeline is designed to handle daily updates efficiently while maintaining consistent results. This consistency is crucial for managing the evolving nature of the dataset and ensuring the accuracy and reliability of the information provided. As new publications are released, they are integrated into the CORD-19 dataset in real-time or near-real-time. This ensures that researchers and users have access to the latest data promptly.

1. **Value:**

This data has proved to be valuable as users of the dataset have leveraged AI-based techniques in information retrieval and natural language processing to extract useful information. The dataset has been downloaded over 200K times in the three months since its release, showing a very positive response to CORD-19. The dataset has been used by clinicians and clinical researchers to conduct systematic reviews, has been leveraged by data scientists and machine learning practitioners to construct search and extraction tools, and is being used as the foundation for several successful shared tasks.

**Expected Relationships and Correlations**

* **Covid-19 Research versus Location:** Certain countries are expected to have more Covid-19 related research than others. This could be due to funding available in first word countries. Also Covid-19 was first seen in certain countries and started spreading out as people travelled to different countries, which could have prompted those countries with earlier Covid-19 cases to do publish research and share information.
* **Hypertension and Severity of Covid-19:** People with existing medical conditions were severely affected by Covid-19 and even hospitalised. Hypertension was on of the medical conditions which worsened sickness in patients who were tested positive for Covid-19. We would expect to see this result in correlation analysis.
* **Gender and Severity of Covid-19:** Gender was also a determinant of the severity of Covid-19 related sickness. Males tend to be more hospitalised than females, however this could be a weak correlation, but still expected either way.

**Conclusion**

The COVID-19 Open Research Dataset (CORD-19) dataset is a valuable resource for studying Covid-19 and using Natural Language Processing. Its substantial volume, variety of data types, high veracity, and historical velocity provide a robust foundation for in-depth analysis. It is a reliable source of research related to Covid-19 as it was sources from reliable organisations such as World Health Organisation. Due to the availability of wider variety of data in this data set such as countries, gender, and text, many different types of informative analysis can be achieved. These includes into looking into geospatial analysis, getting factors that affected Covid-19 related hospitalizations and others.

**Reference**

LuWang, L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Burdick, D., Eide, D., Funk, K., Katsis, Y., Kinney, R., Li, Y., Liu, Z., Merrill, W., Mooney, P., Murdick, D., Rishi, D., Sheehan, J., Shen, Z., Stilson, B., Wade, A. D., Wang, K., Wang, N. X. R., Wilhelm, C., Xie, B., Raymond, D., Weld, D. S., Etzioni, O., & Kohlmeier, S. (2020). CORD-19: The COVID-19 Open Research Dataset. Allen Institute for AI. https://www.semanticscholar.org/cord19