FaCov: COVID-19 Viral News and Rumors Fact-Check Articles Dataset

Shakshi Sharma,[1] **Ekanshi Agrawal**[*],[2] **Rajesh Sharma,**[1] **Anwitaman Datta**[3]

[1]Institute of Computer Science, University of Tartu, Estonia,
[2]Department of Computer Science and Information Systems, BITS Pilani - Hyderabad, India,
[3]School of Computer Science and Engineering, Nanyang Technological University, Singapore
{shakshi.sharma, rajesh.sharma}@ut.ee, f20170233@hyderabad.bits-pilani.ac.in, anwitaman@ntu.edu.sg

Abstract

COVID-19, which was first detected in late 2019 in Wuhan, China, has spread to the rest of the world and is currently deemed a global pandemic. A flux of events triggered by a wide ranging set of factors such as virus mutations and waves of infections, imperfect medical and policy interventions, and vested interest driven political posturing all have created a continuous state of uncertainty and strife. In this verbile environment, misinformation and fake news thrive and propagate easily through the modern efficient all-pervading media and social media tools, resulting in an infodemic running its course in conjunction with the pandemic. In this work, we present a COVID-19 related dataset – FaCov – a compilation of fact-checking articles that examine and evaluate some of the most widely circulated rumors and claims concerning the coronavirus. We have collected articles from 13 very popular fact-checking sources, along with information about the articles and the vetted verity assigned to the claims being evaluated. We also share insights into the dataset to highlight and understand the major conversations and conflicts in narratives encompassing the pandemic.

Introduction

The COVID-19 pandemic that began in the December of 2019[1] has adversely affected society as a whole and in multiple manner, beyond affecting individual's health and physical well-being. The hitherto unfamiliar and often evolving virus, as well as disconcerted societal and political responses to it, have created a hugely uncertain and vitiated environment where fake news and misleading content masqueraded as factual information is produced and spread incessantly. Such false information may be propagated for reasons ranging from fulfilling a vested agenda, to merely for humor and entertainment in the form of satire. Irrespective of the underlying motives, the spread of misinformation regarding disease outbreaks has been a matter of grave concern for a long period of time, a problem large enough to have been given a name of its own – infodemic[2].

The spread of misinformation about the coronavirus has caused a setback in the healthcare community, for example with fear mongering against vaccines and suggestions of unapproved medications, the number of cases of infections has been rising rapidly in several parts of the world, thereby crippling the healthcare infrastructure. As a result, it has become critical to reduce and counter, if not eliminate, the transmission of incorrect information, and to do so, we need a thorough understanding of what constitutes misinformation and the themes they dominantly cover.

Since the outbreak of the pandemic and the resulting infodemic parallelly in cyberspace, academics have started collecting and analyzing misinformation at a scale, as discussed in the *Related Work* section. During our exploration of these works, we noticed an inadequacy of fact-check datasets that contain full-length assessments of prevalent fake news, specifically around COVID-19. We believe that a collection of articles that discuss and fact-check viral news against factual evidence can serve as a one-stop dataset to gain an understanding of the most prominent and influential discussions surrounding COVID-19. This serves as our motivation to build such a dataset.

The earlier studies had a broad emphasis on two research areas. One is to collect and analyze the COVID-19 data in general or on a specific theme (DeVerna et al. 2021; Kim et al. 2021) such as Sinophobic behavior in COVID-19 tweets (Schild et al. 2020), analyzing hate speech and Islamophobic sentiment (Chandra et al. 2021). The other research area is to more broadly detect misinformation associated with COVID-19 and related issues, e.g., masks, vaccines, quarantines, etc. For instance, the authors (Kouzy et al. 2020) study the magnitude of misinformation about COVID-19. In paper (Ferrara 2020), the focus was on the function of bots in disseminating COVID-19 conspiracy theories. The paper (Sharma, Sharma, and Datta 2021) investigates the online discourse on misleading and non-misleading tweets using machine learning and explainability approaches.

[*]Ekanshi Agrawal contributed to this work during her internship at NTU Singapore under the India Connect@NTU Internship program.

[1]https://www.who.int/emergencies/diseases/novel-coronavirus-2019/interactive-timeline#event-0

[2]https://www.who.int/health-topics/infodemic

In this paper, we present FaCov[3] – a collection of full-length articles that discuss and determine the truth value of widespread news and claims around the coronavirus. The articles are collected from 13 prominent fact-checking websites that discuss such news in detail and contain accompanying information such as authors, dates, truth labels, summary, links to the articles, and the full-length fact-check article itself, as discussed in the *Dataset* section. Also discussed in this section, we provide two types of truth classifications to each of the articles: 3-class, with labels like *True, Mix, False*, and a stricter 2-class, with just *True* and *False* labels. In the *Dataset FAIRness* section, we address the FAIR principles of findability, accessibility, interoperability, and reusability of the provided data. In the *Descriptive Analysis* section, we provide additional insights into the dataset by discussing the appearance of various COVID-19 variants in the articles, named entities cited in the dataset, frequently appearing words and phrases, social media platforms mentioned in the articles, and prevalent myths surrounding COVID-19 according to the WHO. Then in the *Discussions* section, we gain a better understanding of the COVID-19 discourse by reviewing articles across the three truth classes.

This dataset spans the time frame of the first two years of the pandemic and addresses limitations of existing data collections on the Covid-19 infodemic in the following manners:

1. We cumulate COVID-19 articles from multiple fact-checking websites, as such websites discuss and validate posts from various social media networks, so diversifying our dataset.
2. Our dataset is rich enough to include a variety of information about COVID-19 and its variants, along with insights such as named entities cited in the dataset, frequently appearing words and phrases, social media platforms mentioned in the articles, and prevalent myths surrounding COVID-19 according to the WHO.

Related Work

This section first discusses different works which have presented various COVID-19 datasets. Then, we look into the literature from the COVID-19 misinformation detection perspective.

COVID-19 Analysis

Researchers look into the COVID-19 infodemic from a variety of perspectives, with a focus on examining COVID-19 discourse on social media platforms. One line of research explored in the research community is to collect and analyze COVID-19 data in general and explore specific themes and topics within it. Examples include analyzing Sinophobic behavior using COVID-19 tweets (Schild et al. 2020), analyzing hate speech, and Islamophobic sentiment in the COVID-19 dataset (Chandra et al. 2021). Social media such as Reddit has been used in the COVID-19 vaccine discussion (Wu, Lyu, and Luo 2021). Specifically, the majority of comments on Reddit threads are about conspiracy ideas. Many researchers explore Twitter social media for COVID-19 data collection and analysis purposes. In a study (Kim et al. 2021), authors released a collection of COVID-19 tweets used to uncover propagation patterns in the network with future implications of their dataset. The paper (DeVerna et al. 2021) collected and analyzed English tweets, along with a dashboard with simple statistics on the COVID-19 vaccine. In another study, over 123 million multilingual tweets were collected (Chen et al. 2020). Utilizing temporal dimension (Malagoli et al. 2021), on a weekly basis, 12 million tweets in two months were released. Moreover, primary themes and user-level engagements around the pandemic were discussed.

Misinformation in COVID-19

The other line of research is detecting misinformation in COVID-19 datasets. Specifically, the authors (Kouzy et al. 2020) study the magnitude of misinformation related to the COVID-19 outbreak. In (Ferrara 2020), the authors focus on the role of bots in spreading conspiracy theories about COVID-19. The online discourse (Sharma, Sharma, and Datta 2021) on misleading and non-misleading COVID-19 vaccine tweets has also been studied utilizing machine learning and explainability approaches. In addition to textual data, a study (Ma and Stahl 2017) employed a multi-modal discourse analysis approach to analyze the textual and visual information within a public anti-vaccine (about vaccines in general, since this work predates Covid-19) Facebook group. The COVID-19 research (Juneja and Mitra 2021) has also focused on retail companies such as Amazon's search and recommendation algorithms, the world's largest retailer. Particularly, the authors undertake two sets of algorithmic audits for vaccination misinformation.

The closest work to ours is the FakeCovid (Shahi and Nandini 2020) data collection. The full-length articles, along with metadata, are collected from 92 fact-checked websites using references from Snopes and Poynter in various languages. However, this dataset lacks authors' information (whereas our dataset has 2956/3088 (i.e., 95.7%) authors' names) which can be utilized to group the authors based on fake and real news. Furthermore, their work lacked important insights while analyzing the dataset, such as coverage of the significant events such as including the discovered COVID-19 variants till December 2021.

Table 1 presents the comparison of our dataset (FaCov) with several of the accessible COVID-19 datasets. The datasets are broadly divided into two types (Column 1). First, data was gathered from social media platforms such as Twitter and then employing fact-checking websites to classify tweets or gather tweet ids. Second, data is gathered directly from fact-checking websites. The criteria to compare datasets is based on several factors. Important ones include whether the authors considered several sources when gathering data (Column 3), whether their dataset contains any author information (Column 4), and whether the data was obtained over a sufficient length of time (column 7). It should be noted that in some cases, researchers have focused only

[3]https://www.doi.org/10.5281/zenodo.5854656

Datasets	Size	Fact Check	Author	Date	Content	Duration
CoAID (Cui and Lee 2020)	926	✓		✓		2019-20
COVID19 Fake News Detection (Patwa et al. 2020)	10,700					-
FibVID (Kim et al. 2021)	1,353	✓	✓	✓		2020
Instagram (Zarei et al. 2020)	5.3K		✓	✓		2020*
COV19Tweets (Lamsal 2021)	310 mn		✓	✓		2020*
COVID-19 Rumor (Cheng et al. 2021)	4,129	✓		✓		2019-21
FakeCovid (Shahi and Nandini 2020)	5,182	✓		✓	✓	2020
COVID-19 misinformation (Memon and Carley 2020)	4,573		✓	✓		2020
FaCov (This Dataset)	3,088	✓	✓	✓	✓	2019-21

Table 1: COVID-19 Datasets. The Datasets column represents the name of the existing COVID-19 datasets and our dataset. The Size column denotes the size of the dataset in terms of the total number of samples (social media posts, claims, articles, etc.). Fact-Check column corresponds to whether there was any use of fact-checking sources in their data collection methodology. The Author and Date columns represent the presence of the author of the fact-check article (or user of the social media post) and the creation date of the fact-check article (or the social media post), respectively. The Content column denotes whether the full-length article/discussion of the news claim from fact-checking websites is present. The duration column indicates the time duration of the collected dataset. The * denotes the ongoing data collection process.

on one source (such as Twitter) for collecting misinformation data. For the author information, we use the term "author" in two different contexts. First, the author refers to the user who publishes posts on social media, while, in fact-checked articles, the author refers to the author who wrote the article.

It should be noted that we gathered articles from the beginning of the pandemic, therefore covering the entire time period. However, one can argue that the collected number of articles is small compared to other datasets. Our main goal is to emphasize the dataset's quality. The dataset's quality could be determined, for instance, by the type of articles we collected and the length of the articles. Specifically, the fact-checking websites consider only those articles which are popularly reported as false by social media users indicating significant articles. Moreover, we could retrieve more information with fact-checked articles that have an average length of 5000 words. In contrast, the maximum length of social media noisy posts, particularly tweets, is just 280 characters long.

To summarize, we present FaCov - a compilation of full-length articles that discuss and determine the truth value of widespread news and claims around the coronavirus. The articles are collected from 13 popular fact-checking websites covering prominent social media platforms. In addition, we have provided insights into the data to ensure that our dataset is rich enough to include the main events, such as COVID-19 variants and popular COVID-19 myths as per WHO. Also, the collection of metadata would aid in detecting and analyzing misinformation.

Dataset

We present the construction of the dataset in this section by going over aspects such as the discovery of sources, collecting articles, and extracting features to generate some insights into the data, which later help in examining the COVID-19 discourse.

Data Collection

Discovering Sources: For the collection of fact-check articles on COVID-19, domains of various fact-checking internet sources were collected from Google Fact Check Tools[4]. This was accomplished by querying terms and phrases such as "coronavirus," "COVID-19," "pandemic," "2019-21 coronavirus pandemic," and their variations, as well as any similar terms provided by the tool. The results comprised of fact-check articles from several websites, and the domains of these websites were noted. This helped in narrowing our search down to those websites that contain relevant news pertaining to COVID-19. We shortlisted 13 such anglophone fact-check websites based on their popularity as well as quality and quantum of relevant content on those sites, and their domains are listed in Table 2, column 1. The reason behind working with these 13 websites was that COVID fact-check articles from these sites were the ones that surfaced in fact-check searches on Google Fact Check Tools. Here, by surface, we mean an appearance in the first 100 search results. This signaled that these websites are popular and carry fact-check articles on widespread claims. Moreover, the Tranco list[5] generated on 30 November 2021 is also used to determine whether or not the websites are genuinely popular. We also looked at a few additional fact-checking websites. We were unable to scrape them, however. As a result, we chose only these 13 websites, and their ranking as per the Tranco list is shown in Table 2.

So, we conclude that the articles from these 13 sites are indeed important and thus, can provide insight into the subjects being addressed in viral social media and news items, as well as in online social discussions on COVID-19.

Collecting Articles: The articles were collected by web-scraping pages from the websites collected earlier, using the Web Scraper[6] browser extension. More specifically,

[4]https://toolbox.google.com/factcheck/explorer

[5]Available at https://tranco-list.eu/list/X74N

[6]https://webscraper.io/

Websites	Rank	w/o Pre	w/ Pre
reuters.com	120	267	116
usatoday.com	177	100	24
indianexpress.com	947	11	5
rappler.com	3044	160	100
afp.com	3210	454	451
politifact.com	3332	1,170	1,169
factcheck.org	5287	735	337
indiatvnews.com	7310	79	25
thelogicalindian.com	53104	1,211	215
boomlive.in	108250	774	474
polygraph.info	244920	449	112
factchecker.in	430705	181	41
covid19factcheck.com	-	19	19
# of articles	-	**5,610**	**3,088**

Table 2: Ranking of the selected websites based on the Tranco list (sorted in descending order). Also, columns w/o Pre and w/ Pre represent number of articles without (before) pre-processing and with (after) pre-processing, respectively.

the sections of these websites that dealt exclusively with COVID-19 related content were scraped. In cases where the website did not have such a specified section, the search functionality within the website was used to query terms related to COVID-19, and the articles in the search results were scraped. Also, in some cases, all articles were scraped, and those unrelated to COVID-19 were filtered out in the pre-processing stage. All the articles collected were then put together into one CSV. The final total number of rows in the dataset is shown in Table 2.

The following information was extracted along with the articles:

- Title of the fact-check article
- URL of the fact-check article
- Claim being discussed in the article (if available)
- Summary of the fact-check article (if available)
- Content of the fact-check article
- Label assigned by the article to the claim
- Author of the fact-check article (if available)
- Date of publication of the article (if available)

Pre-processing and data-cleaning: Pre-processing involved searching the CSV for COVID-related terms and manually reviewing the articles that contained those terms. Irrelevant entries were thus removed, and any entries that contained empty or null values were also eliminated. Furthermore, for articles from some websites (for instance, FactCheck.org), the labels were not explicitly available or accessible for scraping but rather had to be retrieved from the article's content. In such cases, we went over each article's content and labeled them manually. We initially gathered a total of 5,610 fact-check articles and the information that accompanied them, which after the pre-processing totaled **3,088** which became a part of the final dataset, as represented in Table 2. Additionally, Table 3 includes the absolute numbers of non-null values in the dataset. In addition, we found no duplicate claims in the dataset.

Attribute names	Absolute number of the non-null values (out of 3,088)
title	3,088
URL	3,088
claim	2,540
summary	2,286
content	3,088
label	3,088
author	2,956
date	2,929

Table 3: Attribute names with their absolute numbers of non-null values in the FaCov dataset.

Time frame: The fact-check articles in the final dataset cover the period from the start of COVID-19, December 2019, till the first week of December 2021, spanning a period of roughly two years.

Extracted Features

Article Summarization: One of the columns present in the dataset is the *original summary*. However, as this field had 26% null values, we construct the extractive summary (Munot and Govilkar 2014) programmatically using TF-IDF approach, an NLP-based technique (Christian, Agus, and Suhartono 2016). Deep learning-based approaches were also investigated. However, the input size for the majority of the models is limited to 512 words. On average, our *Content* has 5000 words. Other deep learning models that allow for longer text sizes, such as the popular BigBird (Zaheer et al. 2020) transformer model, did not perform well on our dataset compared to the NLP-based technique.

The Term Frequency-Inverse Document Frequency (TF-IDF) is a numerical statistic that indicates how important a term is to a document in a corpus. Its value increases in proportion to the number of times a word appears in a document but is offset by the frequency of the term in the corpus, which helps regulate the fact that certain words are more frequent than others. Following the TF-IDF approach, one required parameter is the number of important sentences to include in the summarization. Thus, we initially compute the mean (μ) and standard deviation (σ) of the number of sentences in each article. Next, we used ($\mu + \sigma$) value as the total number of sentences to include for summarizing every article, which was five. The programmatically generated summary is with the column name as *generated summary*.

Data Annotations: Table 4 refers to scraped websites that have a variety of labels from multiple fact-checking websites such as *False, True, Mostly True, Partly True (Half True), Mostly False, Pants on Fire, Yes, No, Misleading, Altered, Missing Context, Misrepresented,* and *Manipulated.*

Website	True	Mix	False	#
BoomLive	–	Misleading	False	2
FactChecker.in	True	Partly True Misleading	False	4
IndiaTV News	True	Partly True	False	3
Reuters	–	Mostly Fake Half True Misleading	Fake	4
IndianExpress	True	–	False	2
LogicalIndian	True	Half True Mostly True Misleading	Fake, False	6
Covid19 Fact Check	Yes	–	No	2
USAToday	–	Altered Missing Context	False	3
AFPFactCheck	True	Misleading Misrepresented Manipulated	False	5
Politifact	True	Mostly False Half True Mostly True	PoF False	6
FactCheck.org	True	Mostly Fake Half True Mostly True Misleading	Fake	6
Polygraph	True	Likely False Misleading Unclear Unsubstantiated Disputed Uncertain Dangerous	False	9
Rappler	–	Missing Context Altered Photo Altered Video	False	4

Table 4: Truth labels retrieved from the Fact-Check websites after pre-processing under the 3-class labeling, and the number of labels.

We simply aggregate these raw labels into three classes to further analyze data based on the labels and avoid ambiguity in the diversity of labels issued to each article by various fact-checking websites. In addition, two-class labels are included for simplicity, though the three-class labels provide richer context and are used for the analysis of the data presented in this work.

1. 3-class (*True, Mix, False*): In this grouping, the articles that fall under *True, Yes* class are put into *True* class, *False, No, Fake*, and *Pants on Fire (PoF)* articles are put in the *False* class, and the rest of the labels mentioned come under *Mix* class. The *Mix* class label signifies that the statements are somewhat true, but either they lack additional knowledge, context, or important facts, among others. This grouping has been decided as per the labels' definitions in the paper (Shahi and Nandini 2020). The data distribution for the 3-class is as follows: The *True* class has 72 articles, *Mix* has 818 articles, and the *False* class has 2198 articles. Table 4 shows the distribution of labels for each of the 13 websites in the 3-class annotation.
2. 2-class (*True, False*): In this case, we have a *True* class for all the articles labeled as *True* and *Yes*. The rest of the articles fall under the *False* class. The *True* class has 72 articles, whereas the *False* class has 3016 articles.

Dataset FAIRness

FaCov has been made openly available in the CSV format for download in accordance with FAIR standards (Wilkinson et al. 2016). The dataset is **Findable** as it is publicly available on Zenodo, with a digital object identifier (DOI): https://www.doi.org/10.5281/zenodo.5854656. It is also **Accessible** since anyone in the world can retrieve it through the DOI link. The dataset is in the CSV format; hence it is **Interoperable** and can be read by any common CSV reader, spreadsheet program, or programming language. Also, the raw content of the articles is provided directly, allowing the user of our dataset to experiment with and **Reuse** the data for their own studies while making use of the supplementary README file provided to understand the data files in detail.

The *Dataset* section has a detailed discussion of how the dataset and other variables were gathered, as well as a list of metadata items gathered in conjunction with the article. For ease of (re)use and access to articles, we have supplied the URLs of the fact-check articles from the scraped websites to make the original articles available, enhancing reusability, reproducibility, and provenance. While the dataset by itself is self-sufficient, the content and its availability at these URLs are handled by the websites. We have made sure that the selected websites are open to the public and do not have any access restrictions. This is meaningful because if a user of the dataset wants to learn more about an article, they should be able to do so directly from our source. In case an article is later unavailable at the URL, tools like the Internet Archive[7] can also be used for such openly available content.

Descriptive Analysis: What's in the Data?

This section focuses on exploring the FaCov dataset to determine the diversity and representativeness of content captured in it.

Words Appearing on the Axis of the Labels

Visualizing frequently occurring words in the data generally helps acquire a better understanding of the data. In this case, we use the three classes (*True, Mix, False*) to seek the most frequently mentioned words in the *Content* column of the dataset. Specifically, Figure 1 represents the three wordclouds. The frequency of a word in a Figure is proportional to its size. The words that often appear in the *True* label, including *state, variant, data, case, death*. For instance, there are articles about the number of deaths in countries or states

[7]https://archive.org/web/

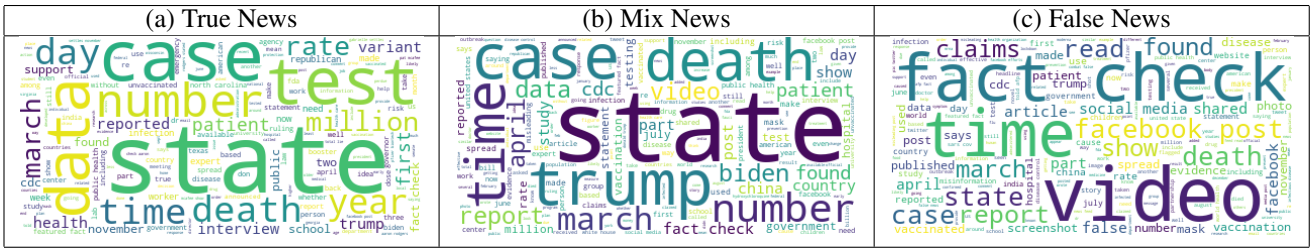

Figure 1: Word Clouds with respect to three labels. The size of the word is proportional to its frequency.

and the rate at which the virus spreads. The words in *Mix* class include *state, time, trump, death, video*. The articles concerned in this class are, for example, the misleading images (videos) posted on social media platforms, Trump posts on healthcare workers, and the number of deaths due to vaccinations. Finally, the *False* class includes *video, time, Facebook, and post* words. This shows that the articles are about the numerous COVID-19 social media posts that are going viral for the wrong reasons and the time duration of the trials and reporting the efficacy of the vaccines.

Frequent Phrases

We identify the common phrases used to produce a brief and appealing *Title* of the articles.

Using the NLTK package[8], we discover the top ten bi-grams and tri-grams (Table 6), sorted in descending order. It can be observed that the bi-grams contain phrases that include labels of the fact-checked articles, social media, preventive measures such as *false claims, social media, face masks, and misleading claims*. While the tri-grams have more generic terms related to claims and social media posts such as *social media post, covid19 fact check, fact check video*.

COVID-19 variants	Title	Content	Total
Alpha	0	26	26
Beta	0	19	19
Gamma	0	11	11
Delta	**12**	**134**	146
Omicron	5	70	75
# of variants	17	260	277

Table 5: The appearance of various COVID-19 variant names. Values indicate the number of articles in which the variant name appeared in the Title and/or Content. The values in bold indicate the highest number of occurrences of a variant in each column, to highlight the most prominent variant being discussed.

COVID-19 Variants

To ensure all the significant events are covered in the dataset, we look for the COVID-19 variants in our dataset listed on the WHO website[9]. In this regard, Table 5 refers to the count of articles that mention the COVID-19 variants. We manually count the total number of variants present per article for each variant in both the *Title* and *Content*. When compared to other variants, the *Delta* variant has the maximum number of articles. The public focus and discussions on variants had echoed the new waves of the pandemic, and the Greek alphabet based nomenclature was also proposed in May 2021[10] following the Delta variant in order to decouple the narrative from the geographic region where a variant is first identified. Omicron was only recently discovered in the period of time captured in our dataset. All these explain and concur with the observed frequencies.

	Bi-grams	Tri-grams
1	covid19, vaccine	social, medium, post
2	fact, check	claim, covid19, vaccine
3	false, claim	post, falsely, claim
4	covid19, death	covid19, fact, check
5	social, medium	fact, check, viral
6	face, mask	covid19, vaccine, contain
7	video, show	make, false, claim
8	bill, gate	covid19, vaccine, cause
9	facebook, post	fact, check, video
10	misleading, claim	claim, circulates, online

Table 6: Top ten Bi-grams and Tri-grams present in the Title column.

Deriving Entity

Named Entity Recognition (NER) aims to find and categorize named entities referenced in the text into pre-defined categories such as people's names, organizations, and so on. The name of the places, people, organization, etc., in the dataset helps in decoding useful information such as geographical locations and celebrities involved in the discussion.

Therefore, we investigate the popularly named entities in our data using the spacy package[11]. Table 7 corresponds to

[8]https://www.nltk.org/api/nltk.html

[9]https://www.who.int/en/activities/tracking-SARS-CoV-2-variants/

[10]https://www.who.int/news/item/31-05-2021-who-announces-simple-easy-to-say-labels-for-sars-cov-2-variants-of-interest-and-concern

[11]https://spacy.io/api/entityrecognizer

NER Tags	Examples
GPE	texas, united states
PERSON	donald trump, joe biden
ORG	congress, democrats
NORP	british, scandinavian
DATE	2020, 2012

Table 7: Top five NER tags with examples Title column in dataset.

the top five named entities extracted from the *Title*, along with examples (keywords) sorted in descending order. In particular, the topmost named entity is Geopolitical Entity (GPE) like *texas, united states* used heavily in our data. The second topmost named entity used is PERSON, which involves the name of the person, such as *donald trump*. The third entity, ORG (name of the organization) such as the *democrats* in the United States. Fourth corresponds to NORP (Nationalities or religious or political groups) includes the person's nationalities, such as *british*. Fifth is the DATE entity, such as *year, month*. Given that our data is from English sources, the dominance of American entities is a natural consequence since they are the most prominently discussed topics in anglosphere cyberspace.

Social Media Platforms

The social media platforms, by reporting false news, aid in the combat against fake news. Including popular social media platforms in the dataset helps cover every important news spread on all platforms. Thus, to ensure that our dataset covers the prominent social media, we look for the social media keywords in our data. Following that, we search for the top social media platforms[12]. However, some social media has not been referred to in our data, such as WeChat. Therefore, we exclude such social media in our analysis. Table 8 represents the references of various social media platforms in both the *Title* and *Content*. We included all social media in the Table mentioned at least once in an article. As obvious, *Facebook* and *Twitter* are the top two most mentioned social media platforms in both *Title* and *Content*. Moreover, we include *Parler* social media as referred to in our dataset but excluded in the top social media list.

COVID-19 Myths

There have been several myths about the COVID-19 epidemic. To explore such information in our dataset, we look into the articles that verify the veracity of these myths. We determined the myths to study based on the myth busters list on the WHO website[13]. Table 9 illustrates the myths we discovered in the *Title*, as well as additional categorization of the myths using the *True, Mix, False* labels. We only include myths that have at least one article addressing them.

[12]https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

[13]https://www.who.int/emergencies/diseases/novel-coronavirus-2019/advice-for-public/myth-busters

Social Media	Title	Content
Facebook	**62**	**2,217**
Twitter	*27*	*1,330*
Instagram	13	482
Weibo	0	19
Youtube	3	337
Whatsapp	18	378
Parler	0	1
Tiktok	3	84
Snapchat	0	3
Pinterest	0	2
Douyin	0	5
Telegram	0	232
# of mentions	**126**	**5,090**

Table 8: References of Social Media in the Title and Content columns. The highest value is bold, while the second topmost value in the specific column is *italicized*.

According to our data, the topmost myth is Hydroxychloroquine which is used to treat malaria and autoimmune diseases and can cure COVID-19. Particularly, the two fact-check articles, that is, *no, ivermectin & Hydroxychloroquine do not treat covid-19, and US medical association did not change stance on Hydroxychloroquine as covid-19 treatment* indicate the myths busted surrounding Hydroxychloroquine. Furthermore, unsurprisingly, the *False* label comprises most of the myths-related articles.

Discussions around COVID-19

In this section, we discuss the key themes surrounding the COVID-19 pandemic present in our dataset. The 3-class labels *True, Mix, and False* discussed in *Dataset* section are used to discover the primary themes (topics). The generative model (non-transformer based models), the LDA[14] approach, has been used to choose the topics. To consider a different approach from LDA, we also looked into a transformer-based model, BerTopic modeling[15]. However, we noticed that the LDA approach yields better results in our case. Particularly, we manually look for the topic name obtained from LDA and BerTopic for all three labels in *Content*. We observed that the LDA topics are generalized and cover the articles' whole agenda more than the BerTopic approach. Next, in LDA, we utilize the Grid Search strategy to investigate the optimal number of topics (themes) in the data. As a result, we identified three significant themes. Furthermore, each primary theme has been investigated further to uncover sub-topics.

True News The first primary theme is *e-learning in Schools*. It has been discussed and fact-checked whether in-person learning helps reduce COVID cases or which states in a certain country have implemented a whole e-learning system. The two sub-topics found under this theme are *Re-*

[14]https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html

[15]https://maartengr.github.io/BERTopic/index.html

Myths about COVID-19, its Prevention, and Cure	# of myths in Title column	Labels		
		True	Mix	False
Ineffectiveness of Alcohol-based hand sanitizers	12	0	4	8
Hydroxychloroquine	**23**	0	8	15
Vitamins & Minerals supplements	7	1	1	5
Usage of Masks while exercising	5	0	2	3
Infection via Water and Swimming	19	0	0	19
Bacteria as a cause	7	0	1	6
Use of Oxygen cylinders and related news	15	0	2	13
Pepper in soup as a cure	1	0	0	1
Spread of infection through houseflies	1	0	1	0
Use of disinfectants on the human body	8	0	1	7
5G networks and relation to COVID	15	1	1	13
Exposure to sun for protection from infection	6	0	1	5
Changes in Life-insurance policies	7	0	2	5
Holding breath as a test for COVID infection	4	0	1	3
Snowy weather as protection	1	0	0	1
Hand dryers as a preventative measure	1	0	0	1
Vaccines against pneumonia as a measure for COVID	3	0	0	3
Rinsing nose with saline water to flush out the infection	7	0	0	7
Consuming garlic	5	0	0	5
Use of antibiotics against the virus	2	0	0	2
# of articles discussing the myths	**149**	**2**	**25**	**122**

Table 9: Fact-checks present in the dataset that are related to widely spread myths according to WHO, extracted from Title column. The myths are then further categorized based on the labels.

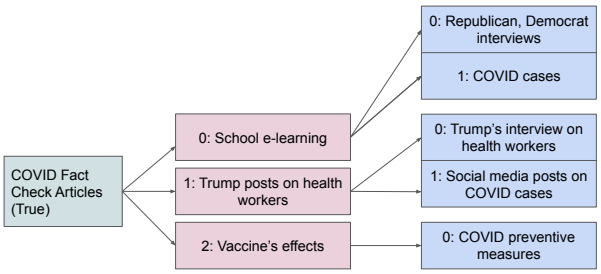

Figure 2: Key themes in True News

publican and Democrats' interviews and COVID cases. Interviews with Republicans and Democrats discuss the economic impact of e-learning during the epidemic. The other theme discusses the graph of the COVID cases after the implementation of e-learning in schools.

The second primary theme is *Donald Trump's social media posts on healthcare workers*. Basically, Trump's message on social media about free vaccination for healthcare employees is accurate and is one of the primary themes in True news. Under this theme, there are two sub-topics; *Trump's interviews on healthcare employees and his social media posts about the amount of COVID-19 cases in various states*. In addition, the same news, along with the description of an increasing number of COVID-19 cases in several states, has been popular in his conversations with health care personnel.

The third primary theme is *Vaccination's effects*. There is a well-known debate and confusion among people about the true risks and effects of vaccination. As a result, there are multiple fact-checked articles concerning vaccination. *COVID prevention measures* are the only sub-topic under this area. With getting vaccinated or not as the main issue, the importance of administering the vaccine to patients as soon as feasible has also been emphasized to prevent and mitigate the COVID-19 pandemic.

Mix News The first primary topic in Mix news is *COVID-19 deaths in states*. Some posts claim to have counted the total number of deaths caused by COVID-19 in different states of a country, which may or may not be accurate but lacks sufficient or further information. As a result, it is always a good idea to double-check such information by studying the full article, as these forms of news might lead us astray by creating a false context. *Misleading posts, photos, and videos* is one sub-topic. As the name suggests, these posts include images or videos in addition to text. For example, photos or videos may be misleading or provide incorrect context in relation to the content. The image could be from an event that occurred a long time ago but is being staged as a recent one.

The second primary topic is *Trump and Biden's statements about China*. There are articles about Trump and Biden regarding China; for example, they talk about Trump's claims that Biden wrote him an apology letter after criticizing travel restrictions on China. However, it was later revealed that Biden never wrote an apology, and Trump put a layer of exaggeration on top of that. Two sub-themes emerged from this theme; *Democrats' debate of COVID reports* and *Trump and Biden's elections in the midst of the*

pandemic. Due to the upcoming elections, both parties became heavily active in discussions of COVID-related problems in order to gain a larger share of the vote.

The third primary topic, *Masks* (common in False news' topic as well). Wearing masks even after being jabbed has been misinterpreted by the majority of people, resulting in the dissemination of fake information. The three sub-topics are; *Misleading posts, pictures, and videos, Drug trials, and Immunity*, and *China's COVID reports and research*. The articles under this theme claims on misleading social media posts on wearing masks, spreading the wrong context on the effectiveness of drug trials, and generating false reports about China's cases and research indicating mask is not neceesary to wear.

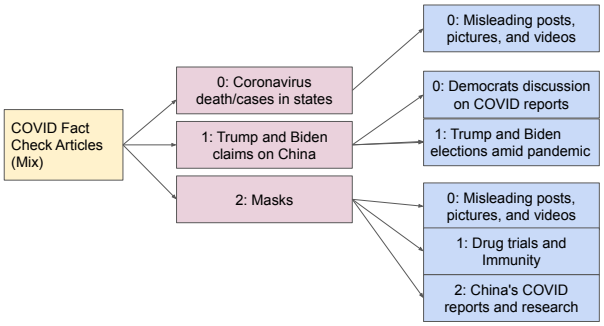

Figure 3: Key themes in Mix News

False News The first primary theme is the *Spread of COVID-19*. There is a lot of fake news out there about how COVID-19 spreads, and people spread incorrect data records on social media. Despite WHO's efforts to dispel COVID-19 falsehoods, people continue to make erroneous claims. This kind of False news has had the greatest impact in all countries. There are three sub-topics under this theme; *False death reports, Masks*, and *False videos on vaccine*. The main sources of propagating the virus include false reports of celebrity deaths on social media claims that wearing masks does not help prevent COVID-19 and posting bogus vaccine videos on social media.

The second primary theme is *Misleading posts, pictures, and videos* (same topic as in Mix news). Social media has been a major source of spreading misleading information, especially in COVID-19 times. Because pictures (images) and videos increase social media visibility in terms of likes and shares, there are posts that contain either false pictures or videos that are irrelevant to the COVID-19 issue. Following the reverse image search, the picture or video was discovered to have occurred in the past on any other occasions. This has been really concerning, and mitigating such misleading information in such a short period of time has proven tough. One sub-theme, *Lockdown news on social media*, falls under this umbrella. The news about the lockdown, such as the length of the lockdown and the directions to follow, is incorrect and lacks any accurate reporting or official links. This emphasizes that we should be cautious of any news until we receive official confirmation.

The third primary theme is *Masks*. Wearing or not wearing masks after vaccination proved to be a substantial point of discussion and the propagation of false news. This frequently discussed *China/Wuhan-related* news quoting that people in China (Wuhan) wore masks even before the COVID-19 outbreak.

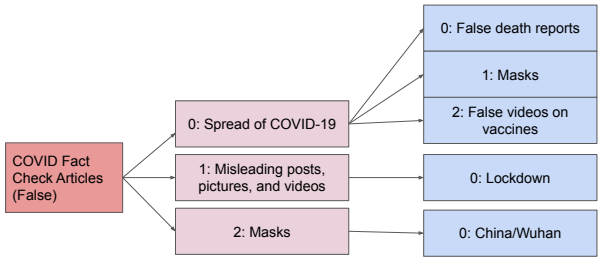

Figure 4: Key themes in False News

Concluding Remarks

Our FaCov dataset, which adheres to the FAIR principles, is a compilation of COVID-19 fact-checking articles from a multitude of sources. Particularly, we retrieved full-length articles as well as metadata in English from 13 fact-checking websites from December 2019 till the first week of December 2021, spanning approximately two years. Searches on Google Fact Check tools were used to identify and shortlist the websites. After performing initial dataset filtering, we annotated the articles into 3-class (*True, Mix, and False*) and 2-class (*True, and False*) categories. This step is necessary to avoid any misinterpretation and harmonize the data, given the variety of labels used by various fact-checking websites. We also boosted the data with an automated summary when the summary was missing in the original source material.

In order to verify that our dataset is rich and diverse enough, we consider many aspects of the COVID-19 pandemic. We analyzed and provided various insights, including covering the articles related to COVID-19 variants to date, articles mentioning common COVID-19 myths as per WHO, most commonly used entities and phrases, and the number of times social media has been referred to in the data. COVID-19 and misinformation might take many different routes in the future. Analyzing the writing styles of False News and True News is one of the possible future directions. We made our dataset accessible, along with the 3-class and 2-class, and generated summaries in accordance with the FAIR principle so that researchers could analyze this dataset further. We anticipate various possible use of the FaCov dataset, for example:

- Automating the detection of health-related misinformation in general.
- Investigate the features that contribute to the detection of health data misinformation and establish explainable frameworks.
- Early identification of articles (and critical users) on social media to prevent the spread of misinformation.

- Help policymakers determine the temporal behavior of misinformation and understand its impacts over time and the shelf-life of genres of misinformation. One can study which of the false claims persists, fades, or is repeated frequently, and which fears, uncertainties, and doubts are exploited to do so? These can, in turn, help policymakers prioritize resources to combat subsets of prominent misinformation.

Acknowledgments

This work has received funding from the EU H2020 program under the SoBigData++ project (grant agreement No. 871042), and by the CHIST-ERA grant CHIST-ERA-19-XAI-010, ETAg (grant No. SLTAT21096).

References

Chandra, M.; Reddy, M.; Sehgal, S.; Gupta, S.; Buduru, A. B.; and Kumaraguru, P. 2021. " A Virus Has No Religion": Analyzing Islamophobia on Twitter During the COVID-19 Outbreak. In *Proceedings of the 32nd ACM Conference on Hypertext and Social Media*, 67–77.

Chen, E.; Lerman, K.; Ferrara, E.; et al. 2020. Tracking social media discourse about the covid-19 pandemic: Development of a public coronavirus twitter data set. *JMIR Public Health and Surveillance*, 6(2): e19273.

Cheng, M.; Wang, S.; Yan, X.; Yang, T.; Wang, W.; Huang, Z.; Xiao, X.; Nazarian, S.; and Bogdan, P. 2021. A COVID-19 Rumor Dataset. *Frontiers in Psychology*, 12.

Christian, H.; Agus, M. P.; and Suhartono, D. 2016. Single document automatic text summarization using term frequency-inverse document frequency (TF-IDF). *ComTech: Computer, Mathematics and Engineering Applications*, 7(4): 285–294.

Cui, L.; and Lee, D. 2020. CoAID: COVID-19 Healthcare Misinformation Dataset. arXiv:2006.00885.

DeVerna, M.; Pierri, F.; Truong, B.; Bollenbacher, J.; Axelrod, D.; Loynes, N.; Torres-Lugo, C.; Yang, K.-C.; Menczer, F.; and Bryden, J. 2021. CoVaxxy: A global collection of English Twitter posts about COVID-19 vaccines. *arXiv e-prints*, arXiv–2101.

Ferrara, E. 2020. # covid-19 on twitter: Bots, conspiracies, and social media activism. *arXiv preprint arXiv: 2004.09531*.

Juneja, P.; and Mitra, T. 2021. Auditing e-commerce platforms for algorithmically curated vaccine misinformation. In *Proceedings of the 2021 chi conference on human factors in computing systems*, 1–27.

Kim, J.; Aum, J.; Lee, S.; Jang, Y.; Park, E.; and Choi, D. 2021. FibVID: Comprehensive fake news diffusion dataset during the COVID-19 period. *Telematics and Informatics*, 64: 101688.

Kouzy, R.; Abi Jaoude, J.; Kraitem, A.; El Alam, M. B.; Karam, B.; Adib, E.; Zarka, J.; Traboulsi, C.; Akl, E. W.; and Baddour, K. 2020. Coronavirus goes viral: quantifying the COVID-19 misinformation epidemic on Twitter. *Cureus*, 12(3).

Lamsal, R. 2021. Design and analysis of a large-scale COVID-19 tweets dataset. *Applied Intelligence*, 51(5): 2790–2804.

Ma, J.; and Stahl, L. 2017. A multimodal critical discourse analysis of anti-vaccination information on Facebook. *Library & Information Science Research*, 39(4): 303–310.

Malagoli, L. G.; Stancioli, J.; Ferreira, C. H.; Vasconcelos, M.; Couto da Silva, A. P.; and Almeida, J. M. 2021. A Look into COVID-19 Vaccination Debate on Twitter. In *13th ACM Web Science Conference 2021*, 225–233.

Memon, S. A.; and Carley, K. M. 2020. Characterizing covid-19 misinformation communities using a novel twitter dataset. *arXiv preprint arXiv:2008.00791*.

Munot, N.; and Govilkar, S. S. 2014. Comparative study of text summarization methods. *International Journal of Computer Applications*, 102(12).

Patwa, P.; Sharma, S.; PYKL, S.; Guptha, V.; Kumari, G.; Akhtar, M. S.; Ekbal, A.; Das, A.; and Chakraborty, T. 2020. Fighting an Infodemic: COVID-19 Fake News Dataset. arXiv:2011.03327.

Schild, L.; Ling, C.; Blackburn, J.; Stringhini, G.; Zhang, Y.; and Zannettou, S. 2020. " go eat a bat, chang!": An early look on the emergence of sinophobic behavior on web communities in the face of covid-19. *arXiv preprint arXiv:2004.04046*.

Shahi, G. K.; and Nandini, D. 2020. FakeCovid–A multilingual cross-domain fact check news dataset for COVID-19. *arXiv preprint arXiv:2006.11343*.

Sharma, S.; Sharma, R.; and Datta, A. 2021. Misleading the Covid-19 vaccination discourse on Twitter: An exploratory study of infodemic around the pandemic. *arXiv preprint arXiv:2108.10735*.

Wilkinson, M. D.; Dumontier, M.; Aalbersberg, I. J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.-W.; da Silva Santos, L. B.; Bourne, P. E.; et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3(1): 1–9.

Wu, W.; Lyu, H.; and Luo, J. 2021. Characterizing Discourse about COVID-19 Vaccines: A Reddit Version of the Pandemic Story. *arXiv preprint arXiv:2101.06321*.

Zaheer, M.; Guruganesh, G.; Dubey, K. A.; Ainslie, J.; Alberti, C.; Ontanon, S.; Pham, P.; Ravula, A.; Wang, Q.; Yang, L.; et al. 2020. Big Bird: Transformers for Longer Sequences. In *NeurIPS*.

Zarei, K.; Farahbakhsh, R.; Crespi, N.; and Tyson, G. 2020. A first instagram dataset on covid-19. *arXiv preprint arXiv:2004.12226*.