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

Shakshi Sharma, Ekanshi Agrawal, Rajesh Sharma, Anwitaman Datta
Institute of Computer Science, University of Tartu, Estonia; Department of Computer Science and Information Systems, BITS Pilani - Hyderabad, India; 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

Introduction

- The COVID-19 pandemic that began in the December of 2019 has adversely affected society as a whole and in multiple manner, beyond affecting individual's health and physical well-being. The spread of misinformation about the coronavirus has caused a setback in the healthcare community.

- *Motivation*:
- Since the outbreak of the pandemic and the resulting infodemic parallelly in cyberspace, academics have started collecting and analyzing misinformation at a scale.
- 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.

Data:

- 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.

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

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**

COVID-19 Datasets in the past (above).

FaCov: Data Collection Process. On the left, ranking of the scraped websites along with number of articles before and after preprocessing the news articles. On the right, the total number of articles in FaCov with their columns details.

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

Descriptive Analysis of FaCov Dataset

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**

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

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

	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

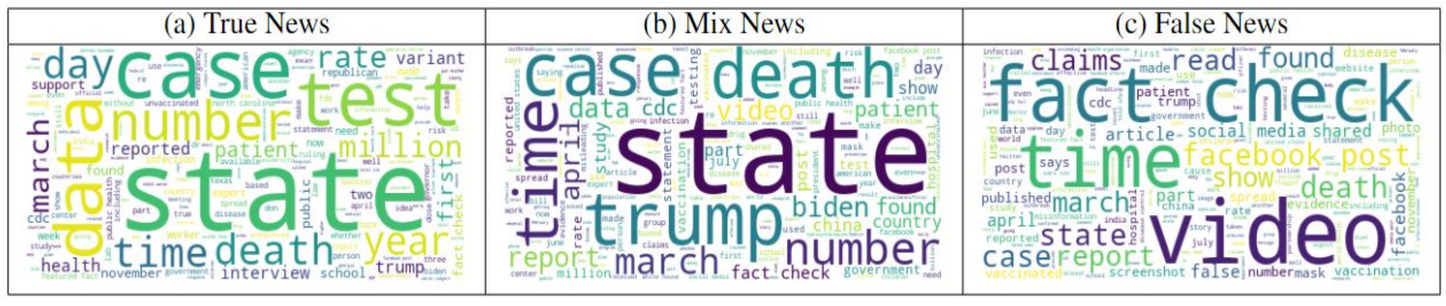
- Named Entity Recognition (NER) aims to find and categorize named entities referenced in the text into pre-defined categories such as people's names, organizations, and so on. The name of the places, people, organization, etc., in the dataset helps in decoding useful information such as geographical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

Findings:

- Delta variant has the maximum number of articles.
- Omicron was only recently discovered in the period of time captured in our dataset.
- 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.

NRFixer: Predicting the fixability of NR bugs

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



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.

Findings:

- The words that often appear in the True label, including state, variant, data, case, death.
- The words in Mix class include state, time, trump, death, video.
- False class includes video, time, Facebook, and post words.

COVID-19 Myths and Future Work

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**

On the left, 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

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.

References:

1. 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.
2. 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.
3. 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: 1016884.
4. Munot, N.; and Govilkar, S. S. 2014. Comparative study of text summarization methods. International Journal of Computer Applications, 102(12).