CoVaxxy: A global collection of English Twitter posts about COVID-19 vaccines

Matthew DeVerna,*1 Francesco Pierri,*1,2 Bao Tran Truong,*1 John Bollenbacher,*1 David Axelrod,1 Niklas Loynes,3,4 Christopher Torres-Lugo,1 Kai-Cheng Yang,1 Filippo Menczer,1,† and John Bryden1

1Observatory on Social Media, Indiana University, Bloomington, Indiana, United States
2Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milano, Italy
3School of Social Sciences, University of Manchester, Manchester, United Kingdom
4Corridor Labs, London, United Kingdom

Abstract

With a large proportion of the population currently hesitant to take the COVID-19 vaccine, it is important that people have access to accurate information. However, there is a large amount of low-credibility information about the vaccines spreading on social media. In this paper, we present a dataset of English-language Twitter posts about COVID-19 vaccines. We show statistics for our dataset regarding the numbers of tweets over time, the hashtags used, and the websites shared. We also demonstrate how we are able to perform analysis of the prevalence over time of high- and low-credibility sources, topic groups of hashtags, and geographical distributions. We have developed a live dashboard to allow people to track hashtag changes over time. The dataset can be used in studies about the impact of online information on COVID-19 vaccine uptake and health outcomes.

Introduction

The COVID-19 pandemic has killed two million people and infected 93 million around the world as of mid-January, 2021 (Dong, Du, and Gardner 2020). Vaccines will be critical in our fight to end the COVID-19 pandemic (Orenstein and Ahmed 2017). It is estimated that around 60-70% of the population will need to be vaccinated against COVID-19 to achieve herd immunity so that virus spread can be effectively suppressed (Aguas et al. 2020). However, recent surveys have found that only 40-60% of American adults reported that they would take a COVID-19 vaccine (Funk and Tyson 2020, Hamel, Kirzinger, and Brodie 2020). With these currently predicted levels of vaccine hesitancy, it is unlikely we will reach herd immunity; COVID-19 will remain endemic in our population.

A possible driver for vaccine hesitancy is the anti-vaccination movement. This movement has been on the rise in the U.S. for two decades, beginning with unfounded fears over a Measles, Mumps and Rubella (MMR) vaccine (Hussain et al. 2018). The vocal online presence of the anti-vaccination movement has undermined confidence in vaccines. Worse, resistance to the COVID-19 vaccines is currently much more prevalent than resistance to the MMR vaccine. Since COVID-19 vaccine hesitancy and its drivers remains understudied, a goal of our project is to help address this gap.

There is a growing body of evidence linking social media and the antivaccination movement to vaccine hesitancy (Broniatowski et al. 2018, Burki 2019, Johnson et al. 2020). Studies show that vaccine hesitancy in one’s peer group is associated with future vaccine refusal (Brunson 2013), and that misinformation spread on social networks is linked to poor compliance with public health guidance about COVID-19 (Roozenbeek et al. 2020). Based on these findings, the core hypothesis behind this project is that the social spread of vaccine misinformation and vaccine hesitancy will impact public health outcomes such as vaccine uptake and COVID mortality rates.

Here we present a collection of English posts related to the COVID-19 vaccines on Twitter. The collection is exempt from IRB review as it only includes tweet IDs of public messages. This allows us to comply with the Twitter Terms of Service while making the data available to both researchers and the general public. Although there has been previous work presenting COVID-19 Twitter datasets (Chen, Lerman, and Ferrara 2020, Huang et al. 2020, Lamsal 2020), our work focuses specifically on discussion of COVID-19 vaccines and related public health outcomes.

The CoVaxxy dataset will enable researchers to study vaccine misinformation and hesitancy, and their relationship to public health outcomes. We will use established techniques to track vaccine misinformation within the data, along with misinformation superspreaders, coordinated campaigns, and automated accounts (Yang, Hui, and Menczer 2019, Yang et al. 2020, Pierri, Piccardi, and Ceri 2020a,b, Pacheco et al. 2020). We will also relate this social media data to geographic public health data (such as COVID-19 mortality and vaccine uptake rates) by using geolocation data within the dataset.

This paper describes relevant aspects of the CoVaxxy dataset, which cover data collection, descriptive analyses of the data and its potential usage, and a live dashboard intended for the public to track key insights drawn from the data. Opportunities and limitations of the dataset are dis-
Dataset curation

Our key data collection goal is to download a complete set of Twitter posts related to COVID-19 vaccines. In this section we describe our methodology for selecting appropriate keywords to achieve such a coverage. We then describe our architecture with server redundancy to maintain an unbroken stream of Twitter data containing these keywords.

Identifying COVID-19 vaccines content

To create as complete a set of Twitter posts related to COVID-19 vaccines as possible, we carefully select a list of keywords through a snowball sampling technique (Conover et al. [2012] Yang, Hui, and Menczer 2019). We start with the two most relevant keywords, i.e., *covid* and *vaccine*, as our initial seeds. Note that keywords also match hashtags, URLs, and substrings. For example, *covid* matches “cnn.com/covid” and “#covid19.” Next, we gather tweets utilizing the filtered stream endpoint of the Twitter API[1] for three hours. From these gathered tweets, we then identify potential keywords that frequently co-occur with the seeds, adding them to our seed list only after manually ensuring they are closely related to our topic. This process was repeated six times between Dec. 15, 2020 and Jan. 2, 2021 with each iteration’s data collection taking place at different times of the day to capture tweets from different geographic areas and demographics. The seed list serves as our initial keyword list.

We further refine the keyword list by manually combining certain keywords into composites, leveraging the query syntax of Twitter’s filtered stream API. For example, using *covid19 pfizer* as a composite matching phrase will capture tweets that contain both “covid19” and “pfizer.” On the other hand, including *covid19* and *pfizer* as separate keywords will capture tweets that contain “covid19” or “pfizer.” Constructing various composites of relevant keywords in this way ensures the dataset is broad enough to include most relevant (English) conversations while excluding tweets that are not related to the vaccine discussion.

Content coverage

To demonstrate the effectiveness of the snowball sampling technique introduced above, we calculate the popularity of each (single or composite) keyword by the number of unique tweets and unique users associated with it.

Figure[1] shows the effect of adding new keywords into the list of streaming filters. The keywords are ranked by popularity. The diminishing growth of popularity suggests that the inclusion of additional keywords is redundant for coverage of users and tweets. The diminishing returns are due to the co-occurrence of multiple keywords and hashtags in a single tweet, especially for the most popular terms. Thus, we believe that our set of keywords provides reasonable coverage and is representative of tweets communicating about COVID-19 vaccines.

[1]https://developer.twitter.com/en/docs/twitter-api/v1/tweets/filter-realtime/overview

Figure 1: Cumulative popularity of keywords in terms of matched tweets (purple, left) and unique users who posted those tweets (green, right), in millions.
As the collection of tweets is intended to persist over time, new keywords will emerge. To ensure that the keyword list remains comprehensive throughout the data collection period, our team will continue to monitor the ongoing public discussion related to COVID-19 vaccinations, should it become necessary to update the list with important emerging keywords.

**CoVaxxy system**

**System architecture**

Our server architecture (Figure 2) is designed to collect and process large quantities of data. The infrastructure is hosted on Jetstream virtual machines (VMs) [Towns et al. 2014; Stewart et al. 2015]. To maintain the integrity of our tweet streaming pipeline, we have incorporated redundancy. We maintain two streamer (stream collection) VMs in different U.S. states so that if one suffers a fault we can use data from the other. These servers connect to Twitter’s filtered stream API to collect tweets that match any of the keywords in real time. We use the language metadata to filter out non-English tweets.

The data from the two streamers is collated on a general purpose server VM where we run data analysis. The server VM is also linked to Anonymous University’s high performance computing infrastructure for running advanced analyses.

We will upload data files to a public data repository on a regular basis. In compliance with Twitter’s Terms, we are only able to share tweet IDs with the public. One can rehydrate the dataset by querying the Twitter API or using tools like Hydrator or twarc.

Finally, a web server provides access to the data on the server VM through applications. An example is the interactive dashboard, described next.

**Dashboard**

Existing COVID-19 visualization tools include those by Johns Hopkins University (Dong, Du, and Gardner 2020) and The Atlantic. These trackers address hospitalization and mortality. Another dashboard from the Fondazione Bruno Kessler covers the infodemic, reporting on the proportions of misinformation and epidemic-related stats (confirmed and death cases) per country. Finally, the Our World in Data COVID-19 vaccination dataset publishes vaccine uptake information by country.

A tool to concurrently explore the relationships between COVID-19 vaccine conversations, vaccine uptake, and epidemic trends is missing. We plan a web-based visualization to fill this void. The CoVaxxy dashboard will track and quantify credible information and misinformation narratives over time, as well as their sources and related popular keywords. The dashboard will focus on the U.S. at the state-level. It will be updated daily. Figure 3 illustrates a simple prototype that displays hashtag sharing behavior at the hourly level. This data will be displayed alongside COVID-19 pandemic and vaccine trends. By highlighting the connection between misinformation and public health actions and outcomes, we hope to encourage the public to be more vigilant about the information they consume on their daily social media feeds in the fight against COVID-19.

**Data characterization**

Our system started to gather tweets on Jan. 4, 2021. Table 1 provides a breakdown of the dataset (as of January 11) in terms of the number of unique users, number of tweets they shared, and numbers of unique hashtags and URLs contained in these tweets.

**Volume**

We show in Figure 4 a time series for the number of tweets collected in our dataset, on an hourly basis. We can notice a decrease in the number of tweets after January 6, which might be driven by the increased media attention surrounding the storming of the U.S. Capitol. In fact, the mean daily number of tweets decreases from 900k tweets in the period of Jan 4–6 to 400k tweets in the period of Jan 7–11.

In Figure 5 we show the distribution of the tweets geolocated in the contiguous United States. We use a naive approach to match tweets to U.S. states: we first extract the user location from the profile (if present) and then match it against a dictionary of U.S. states. Finally, we compute the number of tweets for each state based on the activity of users geo-located in that state. Over 1M users in our dataset have location metadata in their profile; we were able to match approximately 40k users resulting in 600k geo-located tweets. Providing an accurate methodology to geo-locate users is outside the scope of this paper; the reader should consider these results only as an illustration of the insights that can be gained from the CoVaxxy data.

**Hashtags**

Figure 6 lists the most tweeted hashtags in our dataset. We can see that they are largely related to the SARS-CoV-2 vaccine, with one (“#covididiots”) referring to COVID-19 deniers.

Many different conversations can occur concurrently on Twitter, using different hashtags for different topics. To cluster related hashtags, we have grouped them together using a network algorithm. We form a co-occurrence network with

| Users     | Tweets  | Hashtags | URLs    |
|-----------|---------|----------|---------|
| 1,847,067 | 4,768,204 | 39,857   | 983,158 |

Table 1: Breakdown of the data collected between January 3rd and January 10th in terms of unique users, tweets, hashtags and URLs.

[https://github.com/DocNow/hydrator](https://github.com/DocNow/hydrator)
[https://github.com/DocNow/twarc](https://github.com/DocNow/twarc)
[https://covidtracking.com/](https://covidtracking.com/)
[https://covid19lbs.fbk.eu](https://covid19lbs.fbk.eu)
[https://ourworldindata.org/covid-vaccinations](https://ourworldindata.org/covid-vaccinations)

[https://www.nytimes.com/2021/01/06/us/politics/protesters-storm-capitol-hill-building.html](https://www.nytimes.com/2021/01/06/us/politics/protesters-storm-capitol-hill-building.html)
Figure 2: The VM server architecture for the CoVaxxy project. Data flows in the direction of the arrows.

Figure 3: Early prototype of the CoVaxxy web dashboard. This specific visualization allows users to search for and visualize the sharing behavior of specific hashtags within our data set over time.
hashtags as nodes and edges weighted according to how often the linked hashtags co-occur within tweets. Nodes are clustered using the Louvain method (Blondel et al. 2008). Groups with hashtags that are used the most are plotted in Figure 7. We observe groups of hashtags associated with vaccine conspiracy theories ("#greatreset," "#billgates") as well as positive messages ("#stayhome").

**Sources**

In Figure 8 we show the top-10 most shared websites. We exclude “twitter.com,” which accounts for over 3M tweets. These sites are comprised mostly of high-credibility information sources. However, one low-credibility source — “zerohedge.com” — also makes this list (see below for details on the classification). We also observe a large number of links to YouTube, which suggests further investigation will be needed to assess the nature of this shared content.

Figure 9 provides a time series for the prevalence of low- and high-credibility information. We follow an approach widely adopted in the literature (Lazer et al. 2018; Shao et al. 2018; Bovet and Makse 2019; Grinberg et al. 2019; Yang et al. 2020) to label links to news articles based on source reliability. In particular, we use a third-party list of 675 low-credibility sources and 26 hand-selected mainstream sources. Overall, links to low-credibility sources accounts for 24,841 tweets compared to 72,680 tweets linking to our sample of mainstream sources. Readers should note that these numbers do not fully capture the news circulating on Twitter, as the lists we employ cannot be exhaustive.

We further list in Figure 10 the 20 most shared news websites, including both source classes. We notice several unreliable sources (cf. “zerohedge.com” and “bitchute.com”) that exhibit prevalence comparable to more reliable websites.

**Discussion**

In this paper we present a new public dataset tracking discourse about COVID-19 vaccines on Twitter. We characterize the data in several ways, including prominent keywords, geographic distribution of tweets, and clusters of related hashtags. We also present a prototype data dashboard that will visualize statistics and insights from this data.

In future work, we intend to explore the relationship between online discussion of COVID-19 vaccines and public health outcomes, like COVID-19 mortality and vaccine uptake. We will also leverage existing social media analysis tools to track emerging narratives and suspicious accounts, such as bots, coordinated campaigns, and troll farms (Yang, Hu, and Menczer 2019; Yang et al. 2020; Pierri, Piccardi, and Ceri 2020a,b; Pacheco et al. 2020). Finally, we plan to explore models to better understand how vaccine misinformation and anti-vaccine sentiment spreads on social media.

https://iffy.news/iffy-plus/
Figure 7: An overview of the prominent hashtags in the data, clustered into topic groups. A few hashtags characterizing each cluster are shown.

Figure 8: Top-10 sources shared in vaccine-related tweets.

This dataset has a few key limitations. First and critically, Twitter users are not a representative sample of the population, nor are their posts a representative sample of public opinions (Wojick and Hughes 2020). The Twitter filtered stream API also imposes a rate limitation of 1% of all public tweets, which could in future limit our ability to capture all the relevant content. Another potential source of bias is the keyword sampling procedure used to identify and collect COVID-19 vaccine related content, which involved evaluation of keywords to determine what was relevant. Addition-
ally we are unable to fully exclude irrelevant material using only keyword-based content filtering.

The long-term aim of this project is to tackle the ambitious challenge of linking social media observations directly to public health. We hope that researchers will be able to leverage the CoVaxxy dataset to obtain a clearer picture of how vaccine hesitancy and misinformation affect health outcomes. In turn, such insight might enable public health officials to design better strategies for confronting vaccine hesitancy and refusal.

Data supplement
Along with this submission we will include a data file titled covaxxy.tgz. Within that file there is a collection of tweet ids, the list of keywords, and the lists of low- and high-credibility websites. This content will be shared on websites in the future but these will not be anonymous so we can’t list them here.

References

Aguas, R.; Corder, R. M.; King, J. G.; Gonçalves, G.; Ferreira, M. U.; and Gomes, M. G. M. 2020. Herd immunity thresholds for SARS-CoV-2 estimated from unfolding epidemics. medRxiv doi:10.1101/2020.07.23.20160762. URL https://www.medrxiv.org/content/early/2020/11/16/2020.07.23.20160762

Blondel, V. D.; Guillaume, J.-L.; Lambiotte, R.; and Lefebvre, E. 2008. Fast Unfolding of Communities in Large Networks. Journal of Statistical Mechanics: Theory and Experiment 2008(10): P10008. doi:10.1088/1742-5468/2008/10/P10008.

Bovet, A.; and Makse, H. A. 2019. Influence of fake news in Twitter during the 2016 US presidential election. Nature Communications 10(1): 7. ISSN 2041-1723. URL https://doi.org/10.1038/s41467-018-07761-2

Broniatowski, D. A.; Jamison, A. M.; Qi, S.; AlKulaib, L.; Chen, T.; Benton, A.; Quinn, S. C.; and Dredze, M. 2018. Weaponized health communication: Twitter bots and Russian trolls amplify the vaccine debate. American journal of public health 108(10): 1378–1384.

Brunson, E. K. 2013. The Impact of Social Networks on Parents’ Vaccination Decisions. Pediatrics 131(5): e1397–e1404. ISSN 0031-4005, 1098-4275. doi:10.1542/peds.2012-2452.

Burki, T. 2019. Vaccine Misinformation and Social Media. The Lancet Digital Health 1(6): e258–e259. ISSN 2589-7500. doi:10.1016/S2589-7500(19)30136-0. URL https://www.thelancet.com/journals/landig/article/PIIS2589-7500(19)30136-0/abstract

Chen, E.; Lerman, K.; and Ferrara, E. 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.

Conover, M. D.; Gonçalves, B.; Flammini, A.; and Menczer, F. 2012. Partisan asymmetries in online political activity. EPJ Data Science 1(6). doi:https://doi.org/10.1140/epjds6.

Dong, E.; Du, H.; and Gardner, L. 2020. An interactive web-based dashboard to track COVID-19 in real time. The Lancet Infectious Diseases 20(5): 533–534. doi:10.1016/s1473-3099(20)30120-1. URL http://dx.doi.org/10.1016/S1473-3099(20)30120-1

Funk, C.; and Tyson, A. 2020. Intent to get a COVID-19 vaccine rises to 60% as confidence in research and development process increases. Pew Research Center. URL https://www.pewresearch.org/science/2020/12/03/intent-to-get-a-covid-19-vaccine-rises-to-60-as-confidence-in-research-and-development-process-increases/

Grinberg, N.; Joseph, K.; Friedland, L.; Swire-Thompson, B.; and Lazer, D. 2019. Fake news on Twitter during the 2016 U.S. presidential election. Science 363(6425): 374–378. ISSN 0036-8075. URL http://science.sciencemag.org/content/363/6425/374

Hamel, L.; Kirzinger, A.; and Brodie, M. 2020. KFF COVID-19 Vaccine Monitor: December 2020. URL https://www.kff.org/coronavirus-covid-19/report/kff-covid-19-vaccine-monitor-december-2020

Huang, X.; Jamison, A.; Broniatowski, D.; Quinn, S.; and Dredze, M. 2020. Coronavirus Twitter Data: A collection of COVID-19 tweets with automated annotations. Zenodo. doi:10.5281/ZENODO.4393598. URL https://zenodo.org/record/4393598

Hussain, A.; Ali, S.; Ahmed, M.; and Hussain, S. 2018. The Anti-Vaccination Movement: A Regression in Modern Medicine. Cureus 10(7). ISSN 2168-8184. doi: 10.7759/cureus.2919. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6122668/
Johnson, N. F.; Velásquez, N.; Restrepo, N. J.; Leahy, R.; Gabriel, N.; El Oud, S.; Zheng, M.; Manrique, P.; Wuchty, S.; and Lupu, Y. 2020. The Online Competition between Pro- and Anti-Vaccination Views. *Nature* 582(7811): 230–233. doi:10.1038/s41586-020-2281-1.

Lamsal, R. 2020. Coronavirus (COVID-19) Tweets Dataset. IEEE Dataport. doi:10.21227/781w-ef42. URL [https://dx.doi.org/10.21227/781w-ef42](https://dx.doi.org/10.21227/781w-ef42).

Lazer, D.; Baum, M.; Benkler, Y.; Berinsky, A.; Greenhill, K.; et al. 2018. The science of fake news. *Science* 359(6380): 1094–1096. doi:10.1126/science.aao2998.

Orenstein, W. A.; and Ahmed, R. 2017. Simply put: Vaccination saves lives. *Proceedings of the National Academy of Sciences* 114(16): 4031–4033. ISSN 0027-8424. doi:10.1073/pnas.1704507114. URL [https://www.pnas.org/content/114/16/4031](https://www.pnas.org/content/114/16/4031).

Pacheco, D.; Hui, P.-M.; Torres-Lugo, C.; Truong, B. T.; Flammini, A.; and Menczer, F. 2020. Uncovering Coordinated Networks on Social Media. *arXiv preprint arXiv:2001.05658*.

Pierri, F.; Piccardi, C.; and Ceri, S. 2020a. A multi-layer approach to disinformation detection in US and Italian news spreading on Twitter. *EPJ Data Science* 9(35).

Pierri, F.; Piccardi, C.; and Ceri, S. 2020b. Topology comparison of Twitter diffusion networks effectively reveals misleading news. *Scientific Reports* 10: 1372.

Roozenbeek, J.; Schneider, C. R.; Dryhurst, S.; Kerr, J.; Freeman, A. L. J.; Recchia, G.; van der Bles, A. M.; and van der Linden, S. 2020. Susceptibility to misinformation about COVID-19 around the world. *Royal Society Open Science* 7(10): 201199. doi:10.1098/rsos.201199. URL [https://doi.org/10.1098/rsos.201199](https://doi.org/10.1098/rsos.201199).

Shao, C.; Ciampaglia, G. L.; Varol, O.; Yang, K.-C.; Flammini, A.; and Menczer, F. 2018. The spread of low-credibility content by social bots. *Nature Communications* 9: 4787. doi:10.1038/s41467-018-06930-7.

Stewart, C. A.; Cockerill, T. M.; Foster, I.; Hancock, D.; Merchant, N.; Skidmore, E.; Stanzione, D.; Taylor, J.; Tewcke, S.; Turner, G.; Vaughn, M.; and Gaffney, N. I. 2015. Jetstream: A Self-Provisioned, Scalable Science and Engineering Cloud Environment. In *Proceedings of the 2015 XSEDE Conference: Scientific Advancements Enabled by Enhanced Cyberinfrastructure*, XSEDE ’15, 1–8. Association for Computing Machinery. ISBN 978-1-4503-3720-5. doi:10.1145/2792745.2792774. URL [https://doi.org/10.1145/2792745.2792774](https://doi.org/10.1145/2792745.2792774).

Townes, J.; Cockerill, T.; Dahan, M.; Foster, I.; Gaither, K.; Grimshaw, A.; Hazlewood, V.; Lathrop, S.; Lifka, D.; Peterson, G. D.; Roskies, R.; Scott, J. R.; and Wilkins-Diehr, N. 2014. XSEDE: Accelerating Scientific Discovery. *Computing in Science Engineering* 16(5): 62–74. ISSN 1558-366X. doi:10.1109/MCSE.2014.80.

Wojick, S.; and Hughes, A. 2020. Sizing Up Twitter Users. Pew Research Center. URL [https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/](https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/).