Social Network Analysis and Mining (2021) 11:102
https://doi.org/10.1007/s13278-021-00825-0

REVIEW PAPER

An augmented multilingual Twitter dataset for studying the COVID-19 infodemic

Christian E. Lopez[1] · Caleb Gallemore[2]

Received: 15 October 2020 / Revised: 26 September 2021 / Accepted: 28 September 2021 / Published online: 20 October 2021

Abstract

This work presents an openly available dataset to facilitate researchers' exploration and hypothesis testing about the social discourse of the COVID-19 pandemic. The dataset currently consists of over 2.2 billions tweets (count as of September, 2021), from all over the world, in multiple languages. Tweets start from January 22, 2020, when the total cases of reported COVID-19 were below 600 worldwide. The dataset was collected using the Twitter API and by rehydrating tweets from other available datasets, data collection is ongoing as of the time of writing. To facilitate hypothesis testing and exploration of social discourse, the English and Spanish tweets have been augmented with state-of-the-art Twitter Sentiment and Named Entity Recognition algorithms. The dataset and the summary files provided allow researchers to avoid some computationally intensive analyses, facilitating more widespread use of social media data to gain insights on issues such as (mis)information diffusion, semantic networks, sentiments, and the evolution of COVID-19 discussions. In addition, the dataset provides an archive for researchers in the social sciences wishing to have access to a dataset covering the entire duration of the pandemic.

Keywords Twitter · COVID-19 · Named Entity Recognition · Sentiment analysis

1 Introduction

Coronavirus Disease 2019 (COVID-19), is a rapidly spreading illness caused by the Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV2; Khan et al. 2020). On March 11, 2020, the WHO officially classified the COVID-19 outbreak as a global pandemic affecting countries on all inhabited continents (Cucinotta 2020). Since December 2019, when the first cases were reported in Wuhan, China, the number of infected people and fatalities worldwide has increased rapidly (Dong et al. 2020). Both the pandemic itself and policy measures put in place to reduce its spread have had unprecedented economic and social impacts (Nicola et al. 2020), affecting the lives of billions of people. Due to its high infection and death rate alongside its potential for asymptomatic transmission, governments have implemented a wide range of policies to mitigate COVID-19's spread and impact. Such actions began with the Chinese government's order to quarantine Wuhan on January 23rd, 2020, followed relatively quickly by multiple countries declaring states of emergency and implementing strict quarantine and social distancing measures (Nussbaumer-Streit et al. 2020).

Unsurprisingly, COVID-19's spread has been accompanied by a deluge of opinion, commentary, information, and misinformation circulating on social media platforms. Indeed, the social distancing measures required to slow the virus's spread might themselves be encouraging more people to turn to social media to share their experiences.

Infodemiology, or the study of (mis)information's diffusion via digital media, has been a growing concern since the World Wide Web's early years (Eysenbach 2002), but there has been an explosion of literally hundreds of articles on the subject since the pandemic began. These discussions center, in particular, around the concept of an "infodemic," an "overabundance of information—some accurate and some not—that occurs during an epidemic" (Tangcharoensathien et al. 2020). Not only is there a deluge of information from legacy and social media sources, Gazendam et al. (2020) also document exponential growth in medical journal publications—many

✉ Christian E. Lopez
lopezbec@lafayette.edu

[1] Department of Computer Science and Mechanical Engineering Department, Lafayette College, Easton, Pennsylvania, USA

[2] International Affairs Program, Lafayette College, Easton, PA, USA

of them opinion pieces or commentaries—related to the pandemic.

While social media users should by no means be assumed representative of the general public or public opinion (Baumann et al. 2020; Mellon & Prosser 2017), social media is a critical, if also critically flawed, medium of civic discourse (Kruse et al. 2017). Indeed, the infodemiological perspective implies that the dynamics of digital discussions interact with human behavior and the virus's spread in complex ways. Responding to this interest, several social media datasets related to COVID-19 become available at the beginning of the pandemic (e.g., Facebook (Shahi, et al. 2020), news articles (Zhou et al. 2020), Instagram, Reddit (Zarei et al. 2020)).

Our primary interest here, however, is in the numerous Twitter datasets published since the pandemic began. Twitter is a widely used social media platform whose political importance is anchored not only in its millions of daily users but also its (ab)use by elites (Abokhodair et al. 2019; Wells et al. 2020).

While the open availability of numerous Twitter datasets is of great use to researchers, these resources differ in the number, timing, and language of tweets collected, as well as the search keywords used for collection (see Appendix 1). Moreover, while these datasets are a valuable source of text data related to the pandemic, users often still must implement their own Natural Language Processing techniques, which can be computationally intensive, if they are to make meaningful use of this unstructured data. This additional barrier might limit some datasets' utility for interested researchers.

Hoping to facilitate further use of Twitter data for analyzing the COVID-19 infodemic, this work presents a dataset containing tweets collected from all over the world, in multiple languages, starting from January 22nd, 2020. The English and Spanish tweets have been augmented using state-of-the-art Twitter Sentiment and Named Entity Recognition algorithms, providing additional structure for the raw tweet text data and obviating the need to rehydrate tweets from TweetIDs for many research purposes (e.g., sentiment analysis, social media network analysis). In addition to providing metadata at the level of individual tweets, hourly summary statistics of hashtags and mention are provided, which are suitable for semantic network analysis applications. The data collection process and descriptive statistics on the dataset are presented here.

2 Uses of twitter data amid the COVID-19 pandemic

Numerous studies use Twitter data to develop insights related to COVID-19. These analyses range considerably in focus, covering issues such as misinformation, conspiracy theories, and public health surveillance. They further differ substantially in scope. Some studies provide close readings of hundreds of tweets, while other work monitors hundreds of millions (Abdul-Mageed et al. 2020; Larson 2020).

Several studies investigate Twitter data's potential to serve as a tool for public health monitoring amid the pandemic. Al-Garadi et al. (2020), for example, built on Sarker, et al.'s (2020) collection and identification of COVID-19 symptoms, developing a text classifier to monitor tweets for epidemiological purposes. Qin et al. (2020) present a social media search index that might be used to predict numbers of new COVID-19 cases. Mackey et al. (2020), finally, use Twitter data to look for signs of COVID-19 symptoms. Others consider how such data might help assess public responses to pandemic measures. Coftas et al. (2021), for example, use supervised classification to identify positive, negative, and neutral attitudes to COVID-19 vaccines in the first month after the initial successful vaccine announcement. Nurdeni et al. (2021) present a similar analysis of responses to the Sinovac and Pfizer vaccines in Indonesia. The dataset provided here, can facilitate these types of analyses thanks to the Sentiment and Name Entity Recognition algorithms implemented to augment it.

A much larger set of studies, however, uses Twitter data for infodemiological purposes, tracking information and misinformation flows. These studies are too numerous to detail here, but a few prominent examples illustrate the diversity of applications. Gallagher et al. (2020), for example, use a panel of US registered voters' retweeting habits to identify the Twitterverse's authority elites on COVID-19 and the demographic features of their respective followers. Fang and Costas (2020) observe how research on COVID-19 is cited in tweets, while Gilgorić et al. (2020) examine engagement with scientific and governmental authorities. Yang et al. (2020a, b) find links to low-credibility sources to account for more tweeted URLs in March 2020 than links to the CDC, while Pulido et al. (2020) observe in a sample of 1,000 tweets that while false information was tweeted more frequently than true, true information was *re*tweeted more frequently than false. Al-Rawi & Shukla (2020), identify the top 1,000 most active accounts mentioning COVID in a population of approximately 50 million tweets, finding around 12% to be bots, most of which were retweeting news from mainstream outlets, though some also appeared to be boosting survivalist discourse. Yang et al. (2021), similarly, find evidence of some bot activity accelerating shares of links to websites known to post low-credibility content on COVID-19, though verified accounts of these low-credibility outlets, combined with densely connected clusters of accounts retweeting common sources, appear to account for much of the spread. While not all applied to tweets, there are nevertheless burgeoning numbers of misinformation classifiers (Ameur et al. 2021; Elhadad et al.

2020; Malla & Alphonse, 2021; Shaar et al. 2021; Zeng & Chan 2021). All this research notwithstanding, Wicke and Bolognesi (2021) identify a relatively small number of topics in tweets posted between March 20 and July 1, 2020, which, while they respond to major events, occur in remarkably stable proportions across the period. Another subset of studies focuses on prejudice and conspiracy theories linked to the pandemic. Ferrara (2020) studies the role of bots in amplifying COVID-19 conspiracy theories. Vidgen et al. (2020) present a classifier, trained on a 20,000-tweet dataset, to identify anti-Asian prejudice fomented by the pandemic. Tahmasbi, et al. (2020) use text mining to identify the growth and evolution of anti-Chinese hate speech on Reddit and Twitter, and Rodrigues de Andrade et al. (2021) conduct a similar analysis combining sentiment and mentions of China to study popular COVID geopolitics in Brazil. Shahrezaye et al. (2020) investigate conspiracy narratives in a sample of 9.5 million German-language tweets, finding very low rates of both conspiracy narratives and bot activity. Ziem et al.'s (2020) COVID-HATE dataset provides access to egonetworks of accounts with machine classified instances of hate and counterspeech, while Li, Y., et al. (2020) study both stigma and conspiracy theories using manual coding of 7,000 tweets. The combination of sentiment, hashtag, mentions, and name entities in a single large dataset, as presented in this work, facilitates the analysis of social prejudice toward certain populations, individuals, locations or organizations.

Other studies use Twitter to analyze attitudes and emotional responses to the pandemic (Garcia & Berton 2021; Kydros et al. 2021; Tyagi et al. 2021; Venigalla et al. 2020). Abd-Alrazaq et al. (2020), for instance, combine sentiment and topic modeling to identify issues and stances toward them. Yin et al. (2020) combine topic modeling and sentiment analysis to track emotional reactions to different aspects of the pandemic. Jiang et al. (2020) leverage geographic political polarization in the United States to study how political differences affect pandemic debates. Aiello et al. (2020) augment Chen et al. (2020) dataset, using topic modeling and sentiment analysis to study the evolution of English-language debate across the early months of the pandemic according to a model of epidemic psychology. Thelwall and Thelwall (2020) observe differences in word and topic usage by gender.

Dozens of openly available COVID-19-related Twitter datasets exist as of the time of writing (see Appendix 1). Nevertheless, the majority focus on the pandemic's earlier months, generally running from late January or early February to somewhere between March and June 2020. Moreover, most of the available datasets include only Tweet IDs and limited metadata. This means that researchers need to rehydrate the datasets, a process in which users must retrieve twitter data using Tweet IDs. While not technically complicated, rehydration can be time consuming (Chen et al. 2020). Some datasets provide additional structure to the tweets, in the form of sentiment analysis or the results of topic modeling, but only two datasets, one from Feng and Zhou (2020) and one from Gupta et al. (2020), provide both topics and sentiments, which together might allow researchers to bypass rehydration. Gupta et al. (2020) dataset is the largest of the two, at approximately 63 million tweets. Still, this dataset, runs only from late January to the first of July 2020. Hence, it only covers the initial months of the pandemic. Furthermore, while topic modeling might be useful for some researchers, it is a complicated process and will often need to be tailored to their specific interests and needs. Named Entity Recognition (NER), which attempts to identify specific referents, on the other hand, may be more generally applicable for research purposes. However, only one dataset of 8.2 million tweets, created by Dmitrov et al. (2020), features named entity information, and this dataset only covers the period through April 2020.

The dataset we present overcome several limitations of existing datasets. It encompasses a significantly larger corporus of tweets that has been augmented with Sentiment and NER algorithms, as well as with hashtags, mentions, likes, and retweet data. It can help extend these types of studies outlined above by facilitating analysis using named entities, hashtags, mentions, likes, and retweets to identify a relevant corpus without the need to rehydrate voluminous datasets. Further, these data augmentations should allow researchers to perform several social media network analyses and potentially identify clusters of interest (e.g., clusters of misinformation, negative discourse) without the need for rehydration. Because the dataset covers nearly the entire duration of the pandemic to the time of writing, it provides a valuable resource, in particular, for researchers wishing to track changes in topics of discussion, relevant actors, sentiment, and hate speech over time.

3 Dataset

The dataset has over 1.7 billion tweets related to COVID-19 (count as of June, 2021), collected on an ongoing basis and processed with both Sentiment analysis and Named Entity Recognition algorithms. These two operations were selected, in particular, because they are both computationally intensive and provide sufficient data on a given tweet to potentially be useful for future research without further hydration, as shown by previous studies (see Sect. 2). In addition to these data provided at the tweet level, hourly summaries of hashtags, mentions, and the correspondence of hashtags and mentions at the tweet level, suitable for semantic network analysis, are also provided.

3.1 Data collection process

The dataset presented has been continuously collected using the Standard Twitter API since January 22nd, 2020. The tweets are collected using Twitter's trending topics and selected keywords. Some of the keywords used are *virus* and *coronavirus* since 1/22/2020, *ncov19* and *ncov2019* since 2/26/2020, *covid* since 3/22/2020, *rona* since 4/22/2020, *ramadandirumah* (*ramadhan at home* in Bahasa Indonesia), *dirumahaja* (*just (staying) at home* in Bahasa Indonesia), *stayathome* since 5/6/2020, *mask* and *vaccine* since11/18/2020. Moreover, the Twitter dataset from Chen et al. (2020) was used to supplement the dataset presented in this work by hydrating non-duplicated tweets.

As the impact of COVID-19 increased around the world, the research team devoted more computing resources to collecting pandemic-relevant tweets. This is one of the reasons why the number of tweets increased significantly in specific periods (see in Fig. 3). Moreover, given the approach used by Twitter to provide tweets data, it cannot be ensured that the set of tweets in the dataset are a representative sample of all the tweets in a given moment. Users of the data, therefore, should keep in mind the need to normalize the data or select appropriate subsets if conducting temporal analyses.

3.2 Data description

The dataset is organized by hour (UTC) and each hour contains 7 tables: (1) "Summary_Details", (2) "Summary_Hastag", (3) "Summary_Mentions", (4) "Summary_Sentiment", (5) "Summary_NER", (6) "Summary_Sentiment_ES", and (7) "Summary_NER_ES". The description of these seven summary tables is provided in Table 1. For example, given a re-tweet of the original tweet shown in Fig. 1, the information contained on the five tables relevant to this data point is shown in Fig. 2. The "Tweets_ID" feature is used as the primary key to connect all the tables. There is no information about this tweet in the tables with Spanish Sentiment and NER information given the fact this was an English tweet. A detailed description of the features present in each of the tables can be found in the GitHub repository of the dataset [https://github.com/lopezbec/COVID19_Tweets_Dataset].

The English and Spanish tweets are augmented using state-of-the-art Twitter Sentiment and Named Entity Recognition (NER) algorithms. The dataset applies Cliche's (2017) Twitter Sentiment algorithm, which uses an ensemble model of multiple Convolutional Neural Networks and Long Short-Term Memory Networks. According to Otter et al. (2020), it achieves state-of-the-art performance on multiple twitter dataset benchmarks. Comparing several different approaches to sentiment analysis of a sample of Covid-19 tweets, Rustam et al. (2021) find this type of approach to have an accuracy of approximately 75–80%. For each English tweet, the algorithm generates a vector of non-normalized predictions for three sentiment classes: neutral, positive, and negative. Subsequently, the algorithm assigns the tweet to the class with the highest predicted probability. For English-language NER, we used Akbik et al. (2019a, b) algorithms, which takes a pooled contextualized embedding approach. Specifically, we applied the state-of-the-art English NER pre-trained model provided by Akbik et al. (2019b). Similarly, for all the Spanish tweets collected, we applied the Spanish NER pre-trained model provided by Yu et al. (2020). For each English and Spanish tweet, the NER algorithms identify all location (LOC), person (PER), organization (ORG), and miscellaneous (MISC) named entities, as well as the predicted probability for each (i.e., NER_Label Prob). For

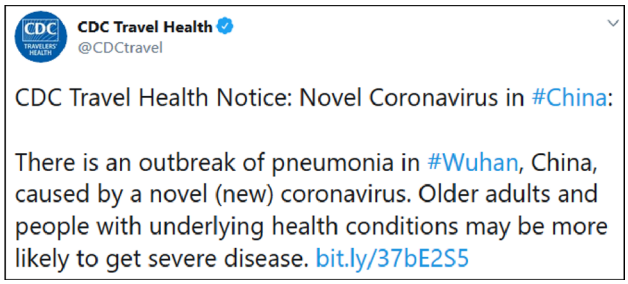

Fig. 1 Example of Tweet Related to COVID-19

Table 1 List and description of dataset tables

Table Name	Description
Summary_Details	This table contains general details about the tweets, such as (i) Tweet ID, (ii) Language, (iii) Geolocation presence or not, (iv) number of likes, (v) number of retweets, (vi) country of tweet (vii) date created
Summary_Hastag	This table contains the Hashtags
Summary_Mentions	This table contains the different mentions
Summary_Sentiment	This table contains the sentiment information of the tweets in English
Summary_NER	This table contains information about the named entity recognized by the NER algorithm of the tweets in English
Summary_Sentiment_ES	This table contains the sentiment information of the tweets in Spanish
Summary_NER_ES	This table contains information about the named entity recognized by the NER algorithm of the tweets in Spanish

1) *2020_01_2_00_Summary_Details Table*

Tweet_ID	Language	Geolocation_coordinate	RT	Likes	Retweets	Country	Date Created
1219772064296361986	en	NO	YES	0	97	NA	Wed Jan 22 00:01:37 +0000 2020

2) *2020_01_2_00_Summary_Hashtag Table*

Tweet_ID	Hashtag
1219772064296361986	#China
1219772064296361986	#Wuhan

3) *2020_01_2_00_Summary_Mentions Table*

Tweet_ID	Mention
1219772064296361986	@CDCtravel

4) *2020_01_2_00_Summary_Sentiment Table*

Tweet_ID	Sentiment_Label	Logits_Neutral	Logits_Positive	Logits_Negative
1219772064296361986	negative	1.573609	-0.9221286	2.314119

5) *2020_01_2_00_Summary_NER Table*

Tweet_ID	NER_Text	Start_Pos	Eng_Pos	NER_Label Prob
1219772064296361986	china	62	67	LOC 0.9999
1219772064296361986	wuhan	107	112	LOC 1.0000
1219772064296361986	china	114	119	LOC 1.0000

Fig. 2 Example of dataset tables

sentiment of Spanish tweets, we used the pre-trained Spanish-language neural network model provided by the Python library sentiment-analysis-spanish 0.0.25. This model predicts the probability of a given Spanish tweet to have a positive sentiment, which is subsequently used to label the tweets as either positive, neural or negative.

The English sentiment algorithm is able to, on average, process a tweet in 0.072 secs using a 2.1 GHz CPU (i.e., 100 million tweets in approximately 83.29 days), but it can easily be parallelized. The NER algorithms, by contrast, cannot be easily parallelized. On average, it can process a tweet in 0.069 secs using a single GeForce RTX 2070 1.62 GHz GPU (i.e., 100 million tweets in approximately 79.83 days). This shows the value of making the sentiment and NER information available to the community since other researchers need not spend the time and computational resources to extract this information.

3.3 Descriptive statistics

The average daily number of tweets collected on the dataset was 151,355.31. The number of tweets collected increased every month from 9,810,850 in January 2020 to its peak of 140,694,770 in August of the same year. Table 2 shows the summary statistics for the daily number of total tweets collected each month until Dec., 2020. Table 2 also shows the daily average and total number of original and re-tweets collected per each month.

Table 3 shows the top five languages present on the dataset. English is the most prominent, accounting for 65.38% of the tweets, followed by Spanish with 13.13%, accounting collectively for nearly four-fifths of our sample. This is the primary reason we decided to augment tweets in these two languages using the Sentiment and NER algorithms. Figure 3 presents the number of tweets from each of the top five languages collected over time.

Information about the number of retweets, likes, and geolocation information was also collected. While there are more than 4 million tweets with geolocation information in the dataset, this represents just 0.23% of the total number of tweets. Figure 4 presents the locations of the tweets with geolocation information since January 22nd, 2020. This figure shows that the most tweets with geolocation information come from the US, Europe, and India.

The sentiment of all the English tweets was estimated using a state-or-the-art Twitter Sentiment algorithm. The dataset contains a total of 513,045,254 English tweets classified as negative (45.3%), 109,730,664 as positive (9.7%), and 510,487,085 as neutral (45.0%). Figure 5 shows the sentiment of all English-language tweets as of May 11th, 2021, while Fig. 6 shows the daily proportion of English tweets given their sentiment (i.e., normalized number of tweets given their sentiment). As shown in Fig. 6, there was an increase in the proportion of negative English tweets in late June 2020, and then a steady decrease by the end of January 2021. These changes in the proportion of negative sentiment may be related to multiple events that occurred in the US during that time frame, like the Black Lives Matter protests and new death forecast from

Table 2 Tweet summary statistics, by month

Month/ year	Avg. OR	Avg. RT	Avg. Total	OR	RT	Total
Jan/2020	5,947.00	30,576.50	35,501.50	1,958,346	7,852,504	9,810,850
Feb/2020	10,978.00	29,918.00	40,604.50	7,624,648	21,944,443	29,568,948
Mar/2020	13,095.50	44,714.50	56,283.00	12,610,824	46,659,589	59,270,412
Apr/2020	30,091.00	89,513.00	119,859.50	20,591,357	60,301,889	80,893,244
May/2020	35,163.00	99,928.50	135,709.00	26,258,213	73,618,083	99,876,289
Jun/2020	51,033.00	142,569.00	193,096.00	34,786,076	95,171,388	129,957,461
Jul/2020	54,131.50	154,737.00	209,566.50	29,441,533	82,903,912	112,345,445
Aug/2020	51,330.50	143,551.00	195,142.00	37,596,182	103,098,588	140,694,770
Sept/2020	50,068	132,040	182,947	35,861,979	92,957,247	128,819,226
Oct/2020	54,489	137,225	198,708	41,062,885	104,195,279	144,962,625
Nov/2020	64,125	111,686	177,062	45,096,171	77,885,575	122,981,746
Dec/2020	64,840	121,149	186,852	49,065,436	87,366,002	133,179,589

Table 3 Distribution of tweets, by language

Language	English	Spanish	Portuguese	French	Bahasa	Others
Number of Tweets	1,133,263,003	227,558,226	77,280,463	50,812,571	44,299,982	200,157,965
Percentage	65.38	13.13	4.46	2.93	2.56	11.55

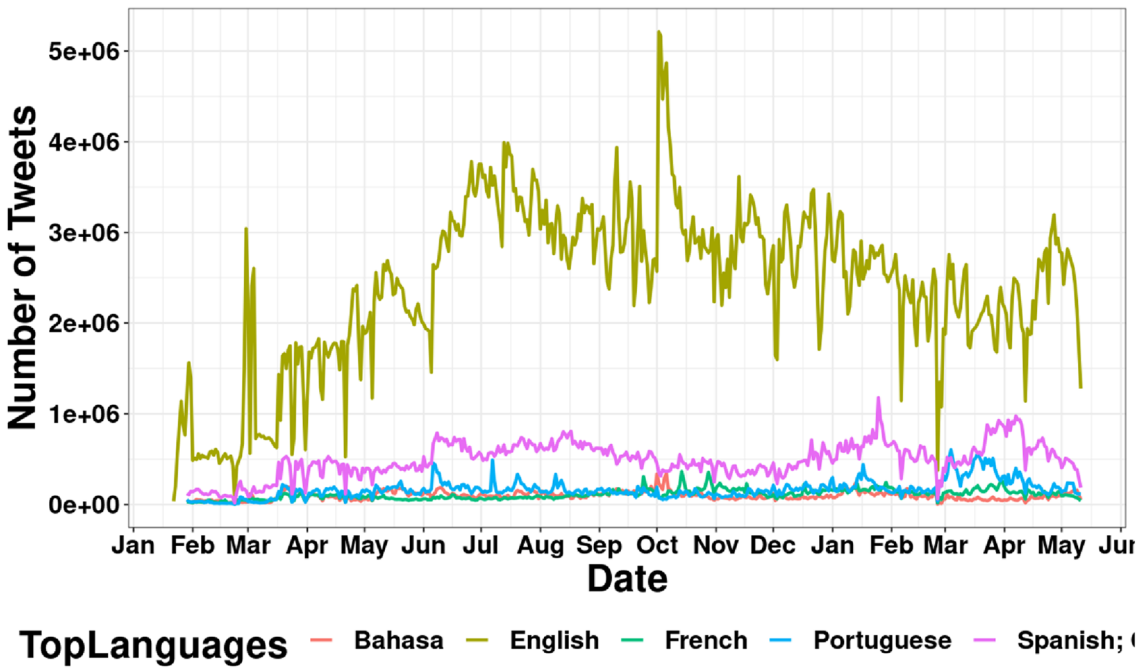

Fig. 3 Tweet frequency across top five observed languages

the CDC in late June 2020, the roll-out of vaccines, and a new US president taking office in late January 2021.

Similarly, the sentiment of all the Spanish tweets was estimated using a Spanish-language sentiment neural network model. The dataset contains a total of 189,137,429 Spanish tweets classified as negative (83.1%), 13,423,158 as positive (5.9%), and 24,997,639 as neutral (11.0%). Figure 7 shows the sentiment of all Spanish tweets, while Fig. 8 shows the daily proportion of tweets in each sentiment category. In comparison with Figs. 7 and 8, it is clear that there are proportionately more negative tweets in the Spanish corpus than in the English.

Lastly, we used a Named Entity Recognition algorithm to extract topics of conversation identified as persons,

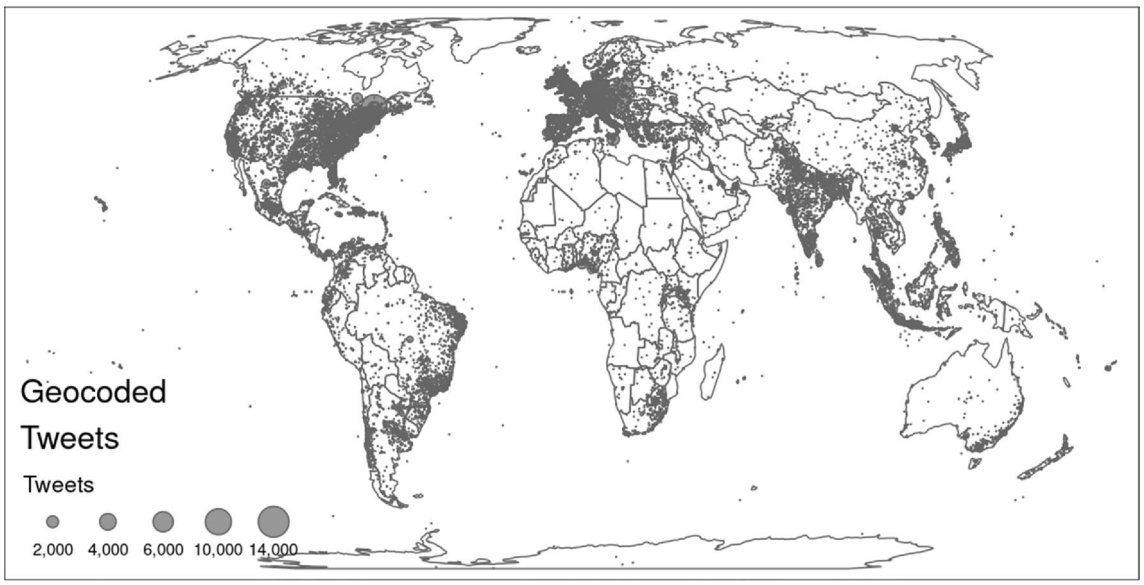

Fig. 4 Map of tweets featuring geolocation information

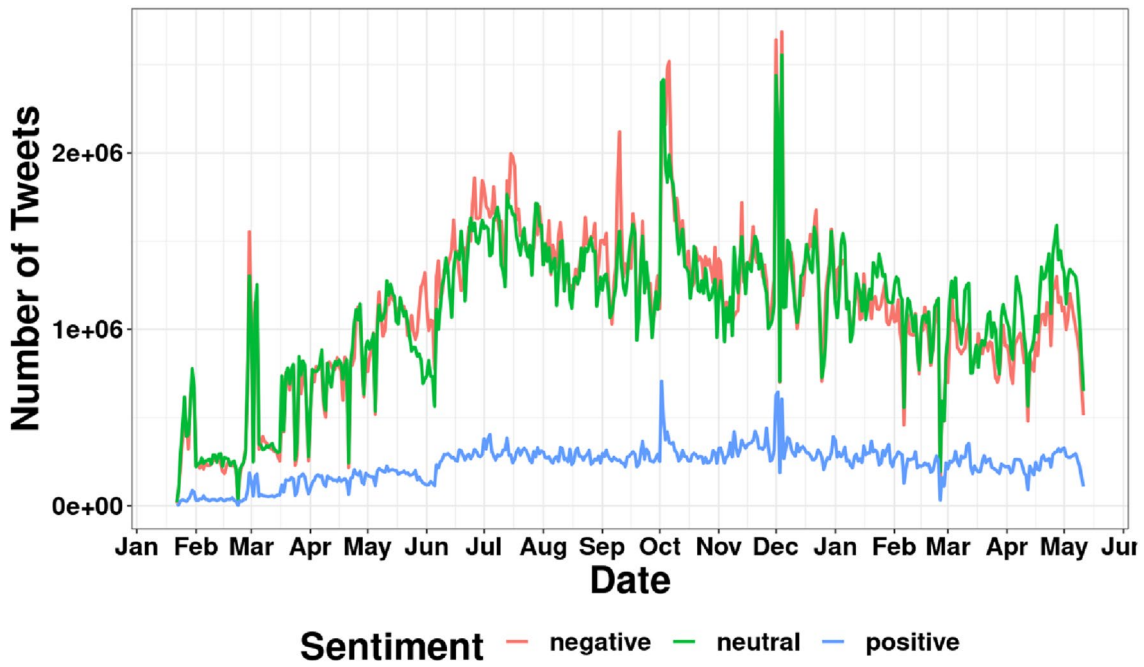

Fig. 5 Sentiment of English-language tweets

locations, organizations, and miscellaneous for both English and Spanish tweets. Table 4 shows the top 5 mentions and hashtags over the entire dataset, as well as the named entities across the dataset of English and Spanish tweets. From Table 4 it can be seen that in some circumstances the Named Entity Recognition algorithm identifies the word "covid" as the subject (i.e., NER Person) a tweet is referring to. Moreover, multiple of the words, mentions, and hashtags could potentially be grouped together given their meaning (e.g., covid19, covi, covi-19). Because we believe it is best to preserve as much of the raw data as possible and leave aggregation decisions up to researchers likely to have diverse potential applications, the dataset does not group these words together.

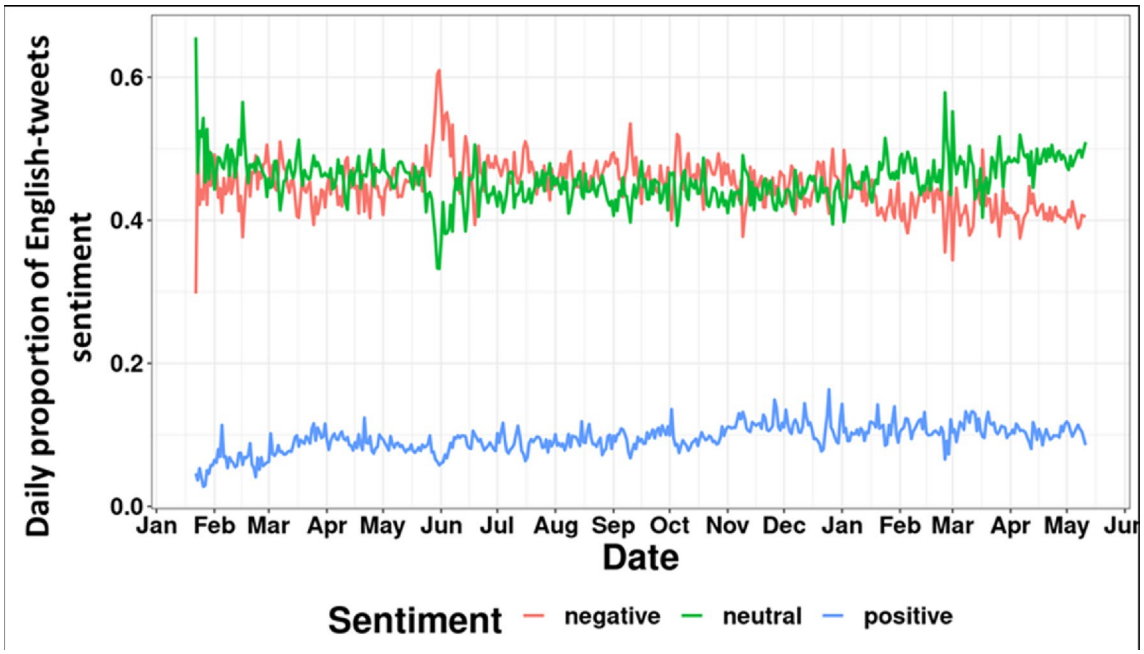

Fig. 6 Daily proportion of English-language tweets sentiment

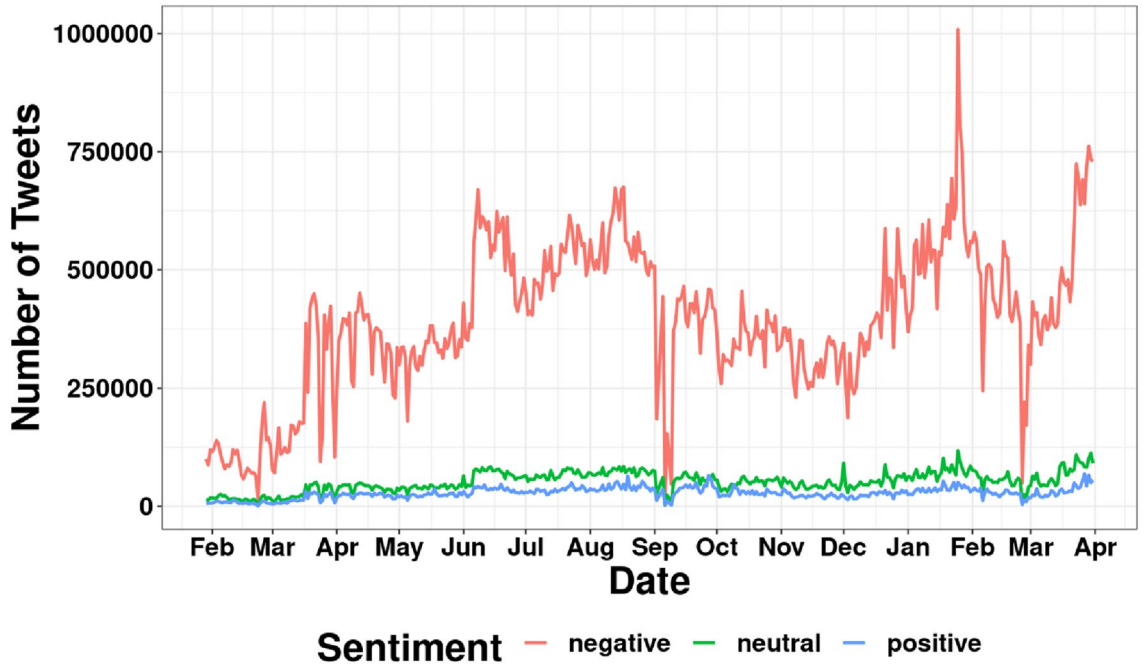

Fig. 7 Sentiment of Spanish tweets

3.4 Data accessibility

The dataset described in this work is available on GitHub at: /lopezbec/COVID19_Tweets_Dataset. This dataset is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International Public License (CC BY-NC-SA 4.0). By their use of this dataset, researchers express their willingness to abide by the stipulations in the license and to remain in compliance with Twitter's Terms of Service. If the user of the dataset would like to obtain all the information provided by the Twitter API, they would need to rehydrate the tweets using the code provided on the GitHub

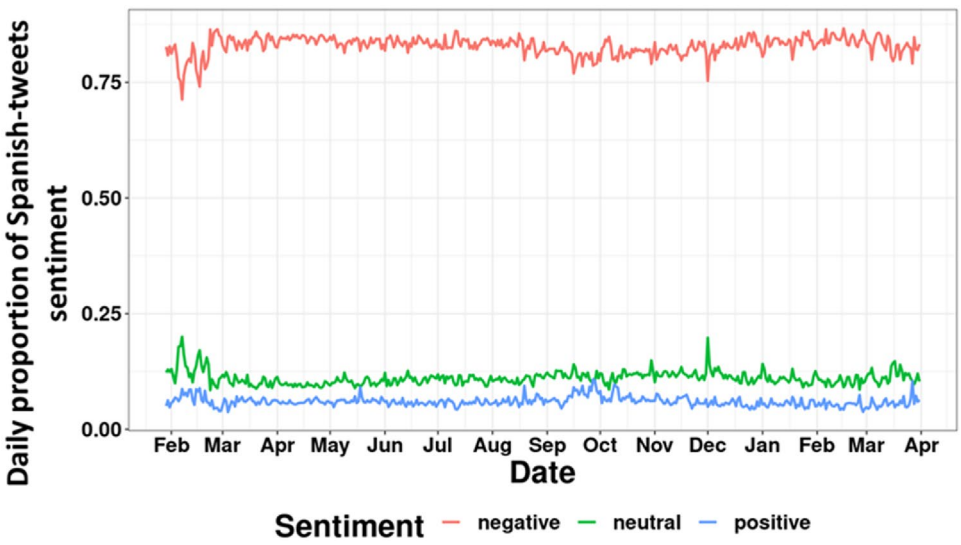

Fig. 8 Daily proportion of Spanish-language tweets by sentiment

Table 4 Top 5 Mentions, hashtags, and named entities

	1st	2nd	3rd	4th	5th
Mentions	@realDonaldTrump	@realdonaldtrump	@mippcivzla	@joebiden	@narendramodi
Hashtag	#covid19	#coronavirus	#covid	#covid-19	#lockdown
NER Person (English/Spanish)	trump/maduro	biden/covid	covid/ivanduque	donal trump/nicolas maduro	fauci/trum
NER Location (English/Spanish)	us/españa	china/italia	uk/china	america/venezuela	india/méxico
NER Organization (English/Spanish)	cdc/gobierno	trump/china	senate/oms	covid/minsaludcol	pfizer/auronplay
NER Miscellaneous (English/Spanish)	covid-19/coronavirus	americans/covid-19	covid/covid19	coronavirus/covid	american/covidã19

repository. This dataset is still being continuously collected and routinely updated.

3.5 Possible use cases

While, as noted above, there already have been other publications providing descriptive analysis of COVID-related tweets, there are, as yet, few studies that move on to hypothesis testing and causal inference (e.g., Gencoglu & Gruber 2020). While numerous researchers might be interested in conducting such analysis, they may not have the expertise or access to computational tools to conduct large-scale Sentiment or Named Entity Recognition analyses. Even the requirement to gather tweets from tweet IDs, known as rehydration, can be an unnecessary barrier to using and reusing large-scale Twitter datasets. By making sentiment analysis and named entity data readily available, this dataset allows researchers to bypass these time- and resource-consuming tasks. Avoiding these barriers will allow researchers to use the dataset as an input into larger analyses. These might include aggregating sentiment data to create a variable for use alongside other data on the pandemic, as done, for example, by Gencoglu and Gruber (2020) or to observe changes in sentiment in response to specific events, as observed by Rodrigues de Andrade et al. (2021) and Tahmasbi et al. (2021). While in these cases, the researchers used data like these to causally model the relationship between disease spread and sentiment (Gencoglu & Gruber 2020) and to identify how major policy events or statements triggered outpourings of anti-Chinese sentiment (Rodrigues de Andrade et al. 2021; Tahmasbi et al. 2021), these are only a few examples of the potential applications of the dataset.

In addition to the possible uses of readily available COVID-19 sentiment data, the named entity data can also

be helpful for researchers, allowing them to track sentiment associated with particular persons, places, or actors related to the pandemic over time, or to identify actors that are generally associated with one another. This can help researchers analyze the emergence and change of conceptual associations between entities over the course of the pandemic, providing a more nuanced picture of how the online discourse on COVID-19 evolves than would be possible by observing sentiment alone. For example, Fig. 9 shows a network with the top 10 most frequent words from hashtag, mentions, and named entities from all the English tweets collected for January 2020. The nodes represent the words. The edges' color represents the average sentiment of the tweets in which both words are present (i.e., red = negative sentiment, black = neutral), while the thickness of the edges represent the frequency in which those words were present together in a tweet. From this figure it can be seen that at the beginning of the pandemic there was a lot of discussion about China and Wuhan, since these nodes have the largest number of edges (i.e., 6 and 4, respectively). Also, it can be shown that the tweets that are related to *China*, *Wuhan*, *Chinese* and have the mention of "*@realDonalTrump*" have the most negative sentiment overall. With the dataset present here, researchers can aggregate common entities (like #Wuhan and wuhan or all the permutations of China) to create more complex semantic networks, analyzing their changes over time to better understand the evolving public sentiment and discourse regarding the pandemic, as well as to find potential clusters of misinformation and high negative sentiment.

To our knowledge, no freely available dataset with sentiment analysis and named entity recognition covers such a long period as the one presented in this work. This makes the dataset potentially useful not only for studying medium-term evolution of online discourse on COVID-19 but also as a historical document of the period. That is, this dataset can serve as an archive for future historians interested in studying an exceptional period in contemporary history.

4 Conclusion and summary

The main objective of this work is to introduce and share with the research community one of the largest openly accessible datasets of tweets with augmented metadata related to the COVID-19 pandemic. The team is continuously collecting and routinely updating the dataset with Sentiment and NER annotations and producing summary files suitable for semantic network and other forms of analysis. The dataset should enable researchers to develop models, test hypotheses, and garner insights from a large archive of Twitter-derived data without the need to rehydrate or conduct computationally prohibitive analyses.

Appendix

(See Table 5).

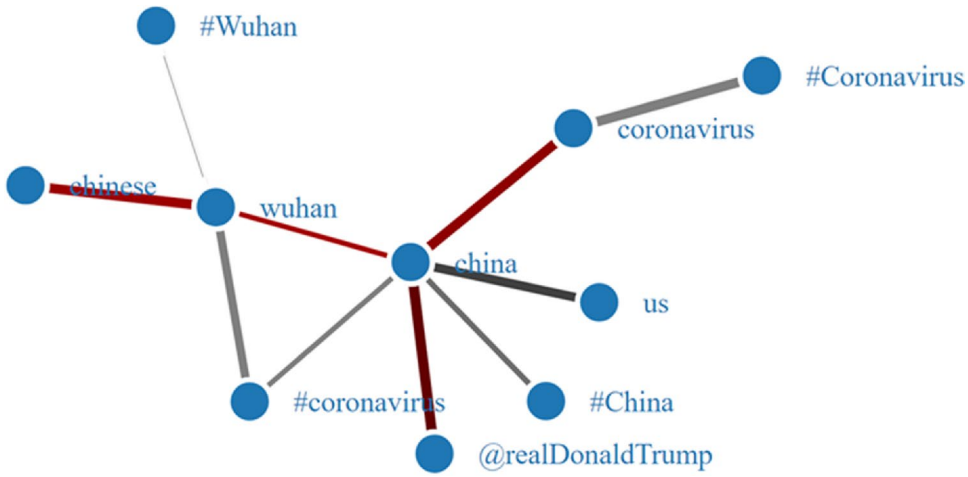

Fig. 9 Network generated from English tweets augmented dataset

Table 5 Openly available COVID-19 Twitter datasets, with features available for download without need for rehydration

Citation	Approximate Tweets	Dates	Tweet ID	Time Stamp	Text	Location	Sentiment	Topic	Other attributes
Abdul-Mageed et al. 2020	1.5×10^9	2007–May 15 2020	✓	✓					
Lamsal 2020	1.3×10^9	Oct 01, 2019–Ongoing	✓						
Banda et al. 2020	1.1×10^9	Jan 27–Ongoing	✓	✓					Hashtag/mention summaries; 1,000 frequent terms
Chen et al. 2020	623×10^6	Jan 28–Ongoing	✓						
Suprem & Pu 2020	600×10^6	Jan 25–Ongoing	✓						
Yang Q et al. 2020	105×10^6	Mar 01–May 15, 2020	✓				✓		
Gupta et al. 2020	63×10^6	Jan 28–July 01,2020	✓	✓	✓	✓	✓	✓	User Metadata; Hashtags; Retweets; Emotions
Ziems et al. 2020	31×10^6	Jan 15–Ongoing	✓	✓		✓		✓	Sampled egonetworks
Gao et al. 2020	25×10^6	Jan 20–Mar 24, 2020	✓	✓					
Dimitrov et al. 2020	8.2×10^6	Oct 2019–Apr 2020	✓	✓			✓		NER; Mentions; Hashtags; User Metadata; URLs
Alqurashi et al. 2020	4.5×10^6	Jan 01– May 30, 2020	✓						
de Melo & Figueiredo 2020	3.9×10^6	Jan–May 2020	✓	✓					Retweets; hashtags
Haouari et al. 2021	2.7×10^6	Jan 27– Jan 31,2021	✓						Propagation network of top 1,000 tweets by day
Feng & Zhou 2020	650×10^3	Jan 25– May 10, 2020	✓	✓		✓	✓	✓	
Sarker et al. 2020	472×10^3	Feb 01–?? Apr 2020							Self-reported COVID-19 symptoms
Cui & Lee 2020	296×10^3	Dec 01 – Nov 01, 2020	✓						User ID; Reply ID; Misinformation detection
Elhadad et al. 2021	110×10^3	Feb 04–Mar 10,2020	✓				✓		Fact- checking annotation
Naseem et al. 2021	90×10^3	Feb–Mar 2020							
Dharawat et al. 2020	61×10^3								Health risk severity
Vidgen et al. 2020	40×10^3	Jan 01 Mar 10, 2020			✓			✓	
Mutlu et al. 2020	14×10^3	Apr 04– Apr 30, 2020	✓						Human-coded stances on Hydroxychloroquine
Ameur et al. 2021	11×10^3	Dec 15, 2019–Dec 15, 2020	✓		✓				Manual topic, misinformation, and negative speech annotations

Table 5 (continued)

Citation	Approximate Tweets	Dates	Tweet ID	Time Stamp	Text	Location	Sentiment	Topic	Other attributes
Memon & Carley, 2020	4.5×10^3	Mar 29; Jun 15/24, 2020	✓	✓				✓	Tweets for users collected in this period

Author' contributions CEL directed the project, composed the code to collect the tweets and to conduct the sentiment analysis and named entity recognition; CG assisted with some data management code, some code for generating summary tables, and some code for data visualization; Both authors collaborated in drafting the manuscript.

Funding This research received no funding, but we are grateful for the support of Jason Simms and Peter Goode for use of Lafayette College's High Performance Cluster.

Availability of data and material Data and additional documentation are available on GitHub at https://github.com/lopezbec/COVID19_Tweets_Dataset

Declarations

Conflicts of interest We have no competing interests to declare.

Code availability Code is available on GitHub at https://github.com/lopezbec/COVID19_Tweets_Dataset

References

Banda JM, Tekumalla R, Wang G, Yu J, Liu T, Ding Y, Artemova K, Tutubalina E, Chowell G. (2020) A large-scale COVID-19 Twitter chatter dataset for open scientific research - An international collaboration. https://zenodo.org/record/4065674#.X38ef9BKjb0

Abd-Alrazaq A, Alhuwail D, Househ M, Hamdi M, Shah Z (2020) Top concerns of tweeters during the COVID-19 pandemic: infoveillance study. *Journal of Medical Internet Research*, 22(4). https://www.jmir.org/2020/4/e19016/

Abdul-Mageed M, Elmandany AR, Pabbi D, Verma K, Lin R (2020) Mega-COV: A billion-scale dataset of 100+ languages for COVID-19. https://arxiv.org/abs/2005.06012

Abokhodair N, Yoo D, McDonald, DW (2015) Dissecting a social botnet: Growth, content and influence in twitter. *18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, 839–851.

Aiello LM, Quercia D, Zhou K, Constantinides M, Šćepanović, S, Joglekar, S (2020) How epidemic psychology works on social media: Evolution of responses to the COVID-19 pandemic. https://arxiv.org/abs/2007.13169

Akbik A, Bergmann T, Vollgraf R (2019) Pooled contextualized embeddings for named entity recognition. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 724–728.

Akbik A, Bergmann T, Blythe D, Rasul K, Schweter S, Vollgraf R (2019) Flair: An easy-to-use framework for state-of-the-art nlp. In Proceedings of 2019. *Conference of the North American Chapter of the Association for Computational Linguistics*, 54–59.

Al-Garadi MA, Yang Y-C, Lakamana S, Sarker, A (2020) A text classification approach for the automatic detection of Twitter posts containing self-reported COVID-19 symptoms. https://openreview.net/pdf?id=xyGSIttHYO

Alqurashi S, Alhindi A, Alanazi E (2020) Large Arabic Twitter dataset on COVID-19. https://arxiv.org/pdf/2004.04315.pdf

Alsudias L, Rayson P (2020) COVID-19 and Arabic Twitter: How can Arab world governments and public health organizations learn from social media? *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*. https://www.aclweb.org/anthology/2020.nlpcovid19-acl.16/

Ameur MSH, Aliane H (2021). AraCOVID19-MFH: Arabic COVID-19 multi-label fake news and hate speech detection dataset. https://arxiv.org/abs/2105.03143

Arora A, Bansal S, Kandpal C, Aswani R, Dwivedi Y (2019) Measuring social media influencer index-insights from Facebook, Twitter and Instagram. J Retail Consum Serv 49:86–101

Baumann F, Lorenz-Spreen P, Sokolov IM, Starnini M (2020) Modeling echo chambers and polarization dynamics in social networks. *Physical Review Letters*, 124: 048301.

Chen E, Lerman K, 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). https://doi.org/10.2196/19273

Cliche M (2017) Bb_twtr at semeval-2017 task 4: Twitter sentiment analysis with cnns and lstms. https://arxiv.org/abs/1704.06125v1

Coftas L-A, Delcea D, Roxin I, Ioanăș C, Gherai DS, Tajariol F (2021). The longest month: Analyzing COVID-19 vaccination opinions dynamics from tweets in the month following the first vaccine announcement. *IEEEAccess*, 9. https://doi.org/10.1109/ACCESS.2021.3059821

Colic N, Furrer L, Rinaldi F (2020) Annotating the pandemic: Named entity recognition and normalisation in COVID-19 literature. https://openreview.net/pdf?id=QbCLrKBvurm

Cucinotta DVM (2020) WHO Declares COVID-19 a Pandemic. Acta Biomed 19(1):157–160

Cui L, Lee D (2020) CoAID: COVID-19 healthcare misinformation dataset. https://arxiv.org/abs/2006.00885

de Melo T, Figueiredo CMS (2020) A first public dataset from Brazilian twitter and news on COVID-19 in Portuguese. Data Brief 32:106179. https://doi.org/10.1016/j.dib.2020.106179

Dharawat AR, Lourentzou I, Morales A, Zhai CX (2020) Drink bleach or do what now? Covid-HeRA: A dataset for risk-informed health decision making in the presence of COVID19 misinformation. https://openreview.net/forum?id=PmY1SNmJlEC

Dimitrov D, Baran E, Fafalios P, Yu R, Zhu X, Zloch M, Dietze S (2020) TweetsCOV19 - A knowledge base of semantically annotated tweets about the COVID-19 pandemic. https://arxiv.org/abs/2006.14492

Dong E, Du H, Gardner L (2020) An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases*, *20*(5), 533–534.

Elhadad MK, Li KF, Gebali F (2020) Detecting misleading information on COVID-19. IEEE Access. https://doi.org/10.1109/ACCESS.2020.3022867

Elhadad MK, Li KF, Gebali F (2021) COVID-19-FAKES: A Twitter (Arabic/English) dataset for detecting misleading information

on COVID-19. In: Barolli L, Li K, Miwa H. (eds) Advances in Intelligent Networking and Collaborative Systems. INCoS 2020. Advances in Intelligent Systems and Computing, vol 1263. Springer, Cham. https://doi.org/10.1007/978-3-030-57796-4_25

Eysenbach G (2002) Infodemiology: the epidemiology of (mis)information. Am J Med 113(9):163–165

Fang Z, & Costas R (2020) Tracking the Twitter attention around the research efforts on the COVID-19 pandemic. https://arxiv.org/abs/2006.05783

Feng Y, Zhou W (2020) Is working from home the new norm? An observational study based on a large geo-tagged COVID-19 Twitter dataset. https://arxiv.org/pdf/2006.08581.pdf

Ferrara E (2020) What types of COVID-19 conspiracies are populated by Twitter bots? *First Monday*, 25(6): http://dx.doi.org/https://doi.org/10.5210/fm.v25i6.10633

Gallagher RJ, Dorshenko L, Shugars S, Lazer D, Welles BF (2020) Sustained online amplification of COVID-19 elites in the United States. https://arxiv.org/abs/2009.07255

Gao Z, Yada S, Wakamiya S, & Aramaki E (2020) NAIST COVID: Multilingual COVID-19 twitter and weibo dataset. https://arxiv.org/abs/2004.08145

Garcia K, Berton L (2021) Topic detection and sentiment analysis in twitter content related to COVID-19 from Brazil and the USA. Appl Soft Comput 101:107057. https://doi.org/10.1016/j.asoc.2020.107057

Gazendam A, Ekhtiari S, Wong E, Madden K, Naji L, Phillips M, Mundi R, Bhandari M (2020) The "infodemic" of journal publication associated with the novel coronavirus disease. J Bone Joint Surg 102(13):e64. https://doi.org/10.2106/JBJS.20.00610

Gencoglu O, Gruber M (2020) Causal modeling of Twitter activity during COVID-19. Computation 8(4):85. https://doi.org/10.3390/computation8040085

Gilgorić K, Ribeiro MH, Müller M, Altunina O, Peyrard M, Salathé M, Colavizza G, West R (2020) Experts and authorities receive disproportionate attention on Twitter during the COVID-19 crisis. https://arxiv.org/abs/2008.08364

Gupta R, Vishwanath A, Yang Y (2020) COVID-19 Twitter dataset with latent topics, sentiments and emotions attributes. https://arxiv.org/abs/2007.06954

Haouari F, Hasanain M, Suwaileh R, Elsayed T (2021). ArCOV-19: The first Arabic COVID-19 Twitter dataset with propagation networks. *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, 82–91. https://www.aclweb.org/anthology/2021.wanlp-1.9/

Jiang J, Chen E, Yan S, Lerman K, Ferrara E (2020) Political polarization drives online conservations about COVID-19 in the United States. *Human Behavior and Emerging Technologies*, 2(3). https://doi.org/10.1002/hbe2.202

Khan S, Siddique R, Shereen MA, Ali A, Liu J, Bai Q, et al. (2020) Emergence of a novel coronavirus, severe acute respiratory syndrome coronavirus 2: biology and therapeutic options. *Journal of Clinical Microbiology*, *58*(5).https://doi.org/10.1128/jcm.00187-20

Kruse LM, Norris DR, Flinchum JR (2017) Social media as a public sphere? Politics on social media. Sociol Q 59(1):62–84

Kydros D, Argyropoulou M, Vrana V (2021) A content and sentiment analysis of Greek tweets during the pandemic. Sustainability 13(11):6150. https://doi.org/10.3390/su13116150

Lamsal R (2020) Design and analysis of a large-scale COVID-19 tweets dataset. Appl Intell. https://doi.org/10.1007/s10489-020-02029-z

Larson HJ (2020) A call to arms: helping family, friends and communities navigate the COVID-19 infodemic. Nature Review Immunology 20:449–450

Li Y, Twersky S, Ignace K, Zhao M, Purandare R, Bennett-Jones B, Weaver SR (2020) Constructing and communicating COVID-19 stigma on Twitter: A content analysis of tweets during the early stage of the COVID-19 outbreak. *International Journal of Environmental Research and Public Health*, 17(18). https://www.mdpi.com/1660-4601/17/18/6847

Mackey T, Purushothaman V, Li J, Shah N, Nali M, Bardier C, Liang B, Cai M, Cuomo R (2020) Machine learning to detect self-reporting of symptoms, testing access, and recovery associated with COVID-19 on twitter: Retrospective big data infoveillance study. *Journal of Medical Internet Research*, 6(2). https://publichealth.jmir.org/2020/2/e19509/

Malla S, Alphonese PJA (2021) COVID-19 outbreak: an ensemble pre-trained deep learning model for detecting informative tweets. Appl Soft Comput 107:107495. https://doi.org/10.1016/j.asoc.2021.107495

Mellon J, Prosser C (2017) Twitter and Facebook are not representative of the general population: Political attitudes and demographics of British social media users. *Research & Politics*, 4(3). https://doi.org/10.1177/2053168017720008

Memon SA, Carley KM (2020) Characterizing COVID-19 misinformation communities using a novel Twitter dataset. https://arxiv.org/pdf/2008.00791.pdf

Mutlu EÇ, Oghaz TA, Jasser J, Tütüncüler E, Rajabi A, Tayebi A, Ozmen O, Garibay I (2020). A stance data set on polarized conversations on Twitter about the efficacy of Hydroxychloroquine as a treatment for COVID-19. https://arxiv.org/abs/2009.01188

Naseem U, Razzak I, Khushi M, Eklund PW, Kim J (2021) COVIDSenti: a large-scale benchmark Twitter data set for COVID-19 sentiment analysis. IEEE Transact Computat Soc Syst. https://doi.org/10.1109/TCSS.2021.3051189

Nicola M, Alsafi Z, Sohrabi C, Kerwan A, Al-Jabir A, Iosifidis C, et al. (2020) The socio-economic implications of the coronavirus pandemic (COVID-19): A review. *International Journal of Surgery*, *78*(185).

Nurdeni DA, Budi I, Santoso AB (2021). Sentiment analysis on Covid19 vaccines in Indonesia: From the perspective of Sinovac and Pfizer. *2021 3rd East Indonesia Conference on Computer and Information Technology*, 9–11 April. https://doi.org/10.1109/EIConCIT50028.2021.9431852

Nussbaumer-Streit B, Mayr V, Dobrescu AI, Chapman A, Persad E, Klerings I, et al. (2020) Quarantine alone or in combination with other public health measures to control COVID-19: a rapid review. *Cochrane Database of Systematic Reviews*, (9).

Otter DW, Medina JR, Kalita JK (2020) A survey of the usages of deep learning for natural language processing. *IEEE Transactions on Neural Networks and Learning Systems.* https://arxiv.org/pdf/1807.10854.pdf

Pulido CM, Villarejo-Carballido B, Redondo-Sama G, Gómez A (2020) COVID-19 infodemic: More retweets for science-based information on coronavirus than for false information. Int Sociol 35(4):377–392

Qin L, Sun Q, Wang Y, Wu K-F, Chen M, Shia B-C, Wu S-Y (2020) Prediction of number of cases of 2019 novel coronavirus (COVID-19) using social media search index. *Environmental Research and Public Health*, 17(7). https://www.mdpi.com/1660-4601/17/7/2365

Rodrigues de Andrade F, Barreto TB, Herrera-Feligreras A, Ugolini A, Lu Y-T (2021) Twitter in Brazil: discourses on China in times of coronavirus. Social Sciences and Humanities Open 3(1):100118. https://doi.org/10.1016/j.ssaho.2021.100118

Rustam F, Khalid M, Aslam W, Rupapara V, Mehmood A, Choi GS (2021) A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis. PLoS ONE 16(2):e0245909. https://doi.org/10.1371/journal.pone.0245909

Shaar S, Alam F, Da San Martino G, Nikolov A, Zaghouani W, Nakov P, Feldman A (2021). Findings of the NLP4IF-2021 shared tasks

on fighting the COVID-19 infodemic and censorship detection. *Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda*, 82–92. http://dx.doi.org/https://doi.org/10.18653/v1/2021.nlp4if-1.12

Shahi GK, Nandini D. (2020). FakeCovid--A Multilingual Cross-domain Fact Check News Dataset for COVID-19. https://arxiv.org/ftp/arxiv/papers/2006/2006.11343.pdf

Shahrezaye M, Meckel M, Steinacker L, et al. (2020) COVID-19's (mis)information ecosystem on Twitter: How partisanship boosts the spread of conspiracy narratives on German speaking Twitter. https://arxiv.org/abs/2009.12905

Shuja J, Alanazi E, Alasmary W, Alashaikh A (2020) COVID-19 open source data sets: a comprehensive survey. Appl Intell. https://doi.org/10.1007/s10489-020-01862-6

Suprem A, Pu C (2020). EDNA-Covid: A large-scale Covid-19 tweets dataset collected with the EDNA streaming toolkit. https://arxiv.org/abs/2010.04084

Tahmasbi F, Schild L, Ling C, Blackburn J, Stringhini G, Zhang Y, Zannettou S (2021). "Go eat a bat, Chang!": On the emergence of sinophobic behavior on web communities in the face of COVID-19. *WWW '21: Proceedings of the Web Conference 2021*, 1122–1133. https://doi.org/10.1145/3442381.3450024

Tangcharoensathien V, Calleja N, Nguyen T, Purnat T, D'Agostino M, et al. (2020). Framework for managing the COVID-19 infodemic: Methods and results of an online, crowdsourced WHO technical consultation. *Journal of Medical Internet Research*, 22(6): https://www.jmir.org/2020/6/e19659/

Thelwall M, Thelwall S. (2020) Covid-19 Tweeting in English: Gender differences. https://arxiv.org/abs/2003.11090

Tyagi P, Goyal N, Gupta T (2021). Analysis of COVID-19 tweets during lockdown phases. *Proceedings of the 9th International Conference on Information and Education Technology*. https://doi.org/10.1109/ICIET51873.2021.9419641

Venigalla ASM, Chimalakonda S, Vagavolu D (2020). Mood of India during Covid-19 - An interactive web portal based on emotion analysis of Twitter data. CSCW '20 Companion: Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing, 65–68. https://doi.org/10.1145/3406865.3418567

Vidgen B, Botelho A, Broniatowski D, Guest E, et al. (2020). Detecting East Asian prejudice on social media. https://arxiv.org/abs/2005.03909

Wells C, Shah D, Lukito J, Pelled A, Pevehouse JC, Yang J (2020) Trump, Twitter, and news media responsiveness: a media systems approach. New Media Soc 22(4):659–682

Wicke P, Bolognesi P (2021) Covid-19 discourse on Twitter: How the topics, sentiments, subjectivity, and figurative frames changed over time. Frontiers in Communic. https://doi.org/10.3389/fcomm.2021.651997

Yang K-C, Torres-Lugo C, Menczer F (2020) Prevalence of low-credibility information on Twitter during the COVID-19 outbreak. https://arxiv.org/abs/2004.14484

Yang Q, Alamro H, Albaradei S, Salhi A, Lv X, et al. (2020) SenWave: Monitoring the global sentiments under the COVID-19 pandemic. https://arxiv.org/abs/2006.10842

Yang K-C, Pierri F, Hui P-M, Axelrod D, Torres-Lugo C, Bryden J, Menczer F (2021) The COVID-19 infodemic: twitter versus facebook. Big Data and Society, January-June. https://doi.org/10.1177/20539517211013861

Yin H, Yang S, Li J (2020) Detecting topic and sentiment dynamics due to COVID-19 pandemic using social media. https://arxiv.org/abs/2007.02304

Yu J, Bohnet B, Poesio M (2020). Named entity recognition as dependency parsing. https://arxiv.org/abs/2005.07150

Zarei K, Farahbakhsh R, Crespi N, Tyson G. (2020). A first Instagram dataset on COVID-19. https://arxiv.org/abs/2004.12226

Zeng J, Chan C-h (2021). A cross-national diagnosis of infodemics: Comparing the topical and temporal features of misinformation around COVID-19 in China, India, the US, Germany and France. *Online Information Review*. https://www.emerald.com/insight/content/doi/https://doi.org/10.1108/OIR-09-2020-0417/full/html

Zhou X, Mulay A, Ferrara E, Zafarani R (2020) ReCOVery: A multimodal repository for COVID-19 news credibility research. https://arxiv.org/abs/2006.05557

Ziems C, He B, Soni S, Kumar S. (2020) Racism is a virus: Anti-Asian hate and counterhate in social media during the COVID-19 crisis. https://arxiv.org/abs/2005.12423