An Augmented Multilingual Twitter Dataset for Studying the COVID-19 Infodemic

Christian E. Lopez (lopezbec@lafayette.edu)
Lafayette College https://orcid.org/0000-0003-2801-4618
Caleb Gallemore
Lafayette College https://orcid.org/0000-0003-1703-0241

Data Note

Keywords: Twitter, COVID-19, named entity recognition, sentiment analysis, dataset

Posted Date: October 23rd, 2020

DOI: https://doi.org/10.21203/rs.3.rs-95721/v1

An Augmented Multilingual Twitter Dataset for Studying the COVID-19 Infodemic

Christian E. Lopez[1, 2]
Corresponding Author
Email: lopezbec@lafayette.edu
ORCID ID: 0000-0003-2801-4618
[1]Computer Science Department, Lafayette College, Easton, PA 18042
[2] Mechanical Engineering Department, Lafayette College, Easton, PA 18042

Caleb Gallemore[3]
ORCID ID: 0000-0003-1703-0241
[3]International Affairs Program, Lafayette College, Easton, PA 18042

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.

Conflicts of interest/Competing interests: We have no competing interests to declare.

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

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

Authors' 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.

An Augmented Multilingual Twitter Dataset for Studying the COVID-19 Infodemic

Abstract

We present an openly available dataset to facilitate researchers' exploration of popular discourse about the COVID-19 pandemic. The dataset, whose collection is ongoing, currently consists of over 780 million tweets, from all over the world, in multiple languages. Tweets start from 22 January 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 another openly available database. To facilitate access for other researchers, the English-language tweet data has been augmented by state-of-the-art Twitter sentiment and named entity recognition algorithms. The dataset and the summary files we provide 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, sentiment, and the evolution of COVID-19 discussions. The insights extracted from such analyses could help inform policy and advocacy work amid the current and future pandemics.

Keywords: Twitter; COVID-19; named entity recognition; sentiment analysis, dataset

1. Introduction

Coronavirus Disease 2019 (COVID-19), is a rapidly spreading disease 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 of COVID-19 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, to, most recently, 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 going 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 of them opinion pieces or commentaries - related to the pandemic.

While social media users should by no means be assumed to be representative of the general public or public opinion (Baumann, et al. 2020; Mellon & Prosser 2017), social media is a critical, and also a 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 have become available in just a few months (e.g., Facebook (Shahi, et al.i 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 solidified not only in its millions of daily users but also its (ab)use by elites (e.g., 45th President of the USA; 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 extract valuable insights from 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, we present a dataset containing tweets collected from all over the world, in multiple languages, starting from January 22nd, 2020. The dataset has 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. In addition to providing metadata at the level of individual tweets, we also provide hourly summary statistics of hashtags and mentions, suitable for semantic network analysis applications. This research note describes the data collection process and presents descriptive statistics on the dataset.

2. Uses of Twitter 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 investigating hundreds of tweets, while other work monitors hundreds of millions (Abdul-Mageed, et al. 2020; Larson 2020).

Several studies investigate the 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.

A much larger set of studies, however, uses Twitter data for infodemiological purposes, information and misinformation flows. These studies, many in the preprint phase, 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. (2020) find links to low-credibility sources to account for more tweet 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), finally, 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.

A specific subset of these studies focuses on prejudice and conspiracy theories linked to the pandemic. Ferrara (2020), for example, uses Chen, et al.'s (2020) dataset to study 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. 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.

Other studies use Twitter as a means to analyze attitudes and emotional responses to the pandemic. 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. Working with the dataset reported in Chen et al. (2020), Jiang, et al. (2020) leverage geographic political polarization in the United States to study how political differences affect pandemic debates. Aiello, et al. (2020) also augment Chen, et al.'s (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 & Thelwall (2020) observe differences in word and topic usage by gender.

In total, we have found a remarkable number (20) of openly available COVID-19-related Twitter datasets as of the time of writing, most not yet peer-reviewed (see Appendix 1). Some are the products of ongoing data collection efforts, but the majority focus on the pandemic's earlier months, generally running from late January or early February to somewhere between March and June 2020. Trying to keep on the right side of Twitter's somewhat ambiguous terms of use restrictions, most of the available datasets include only Tweet IDs and limited metadata, requiring rehydration, a process of retrieving tweet data using Tweet IDs that, while not technically complicated, 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, 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.'s (2020) dataset is the largest of the two, at approximately 63 million tweets, running from late January to the first of July 2020. Furthermore, while topic modeling might be useful for some researchers, it is a complicated process and will often need to be tailored to researchers' specific interests and needs. Named entity recognition (NER), which attempts to identify specific referents, may be more generally applicable for research purposes. However, only one dataset of 8.2 million tweets, created by Dmitrov, et al. (2020), includes named entity information, and this dataset only covers the period through April 2020.

3. Dataset

We present a dataset of over 785,118,723 tweets related to COVID-19 (count as of Sept. 19st, 2020), collected on an ongoing basis and processed with both sentiment analysis and named entity recognition. We select these two operations, 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. In addition to these data provided at the tweet level, we provide hourly summaries of hashtags, mentions, retweets, likes, and the copresence of hashtags and mentions at the tweet level, suitable for semantic network analysis.

3.1. Data Collection Process

The dataset presented has been continuously collected using the Standard Twitter API since January 22nd, 2020. As of September 19st, 2020 there were a total of 785,118,723 tweets collected. 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), *dirumahaja* (*just (staying) at home* in Bahasa), *stayathome* since 5/6/2020. Moreover, the Twitter dataset from Chen et al. (2020) was used to supplement the dataset presented in this work by hydrating non-duplicated tweets. While the dataset contains tweets from 64 languages, only English-language tweets were collected from 22 January to 28 January 2020.

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. 5). Users of the data, therefore, should keep in mind the need to normalize the data or select appropriate subsets of the data if conducting temporal analyses.

3.2. Data Description

The dataset is organized by hour (UTC) and each hour contains five tables: (1) "Summary_Details", (2) "Summary_Hastag", (3) "Summary_Mentions", (4) "Summary_Sentiment", and (5) "Summary_NER". The description of the features 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.

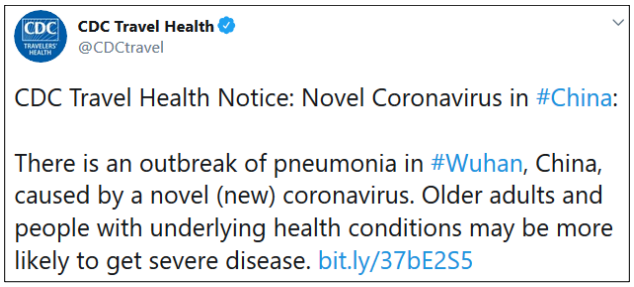

Figure 1. Example of Tweet Related to COVID-19

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

Figure 2. Example of dataset tables

Table 1. List and Description of dataset features

Feature Name	**Description**
Tweet_ID	Integer representation of the tweet's unique identifier*
Language	When present, indicates a BCP47 language identifier corresponding to the machine-detected language of the Tweet text *
Geolocation_cordinate	Indicates whether or not the geographic location of the tweet was reported*
RT	Indicates if the tweet is a retweet (YES) or original tweet (NO)
Likes	Number of likes for the tweet
Retweets	Number of times the tweet was retweeted
Country	When present, indicates a list of uppercase two-letter country codes from which the tweet comes*
Date_Created	UTC date and time the tweet was created
Hashtag	Hashtag (#) present in the tweet
Mention	Mention (@) present in the tweet
Sentiment_Label	Most probable tweet sentiment (neutral, positive, negative) §
Logits_Neutral	Non-normalized prediction for neutral sentiment §
Logits_Positive	Non-normalized prediction for positive sentiment§
Logits_Negative	Non-normalized prediction for negative sentiment§
NER_text	Text stating a named entity recognized by the NER algorithm†
Start_Pos	Initial character position within the tweet of the *NER_text*†
End_Pos	End character position within the tweet of the *NER_text*†
NER_Label Prob	Label and probability of the named entity recognized by the NER algorithm†

* *See Twitter API documentation for more information*

§*See Sentiment algorithm details below*

†*See NER algorithm details below*

The English tweets collected using the Twitter API are augmented using state-of-the-art Twitter sentiment and named entity recognition (NER) algorithms. We use Cliche's (2017) Twitter Sentiment algorithm for sentiment analysis. The algorithm 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. 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 (i.e., *Sentiment Label*). For NER, we use Akbik, et al.'s (2019a) algorithm, which takes a pooled contextualized embedding approach. Specifically, the state-of-the-art English NER pre-trained model provided by Akbik et al. (2019b) is used in this work. For each English tweet, the algorithm identifies 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*).

The 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 algorithm, 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).

3.3. Descriptive Statistics

The average daily number of tweets collected on the dataset was 130,484.61. The number of tweets collected increased every month from 9,810,850 in January to 140,694,770 by the end of August. Table 2 shows the summary statistics for the daily number of total tweets collected each month until Sept. 19st, 2020. Table 2 also shows the daily average and the total number of original and retweets collected per month.

Table 2. Tweet summary statistics, by month.

Month	Avg. OR	Avg. RT	Avg. Total	OR	RT	Total
Jan	5,947.00	30,576.50	35,501.50	1,958,346	7,852,504	9,810,850
Feb	10,978.00	29,918.00	40,604.50	7,624,648	21,944,443	29,568,948
Mar	13,095.50	44,714.50	56,283.00	12,610,824	46,659,589	59,270,412
Apr	30,091.00	89,513.00	119,859.50	20,591,357	60,301,889	80,893,244
May	35,163.00	99,928.50	135,709.00	26,258,213	73,618,083	99,876,289
Jun	51,033.00	142,569.00	193,096.00	34,786,076	95,171,388	129,957,461
Jul	54,131.50	154,737.00	209,566.50	29,441,533	82,903,912	112,345,445
Aug	51,330.50	143,551.00	195,142.00	37,596,182	103,098,588	140,694,770
Sept*	50,472	137,207	188,427	22,884,717	60,674,673	83,559,390

**Until Sept. 19th, 2020*

Table 3 shows the top 5 languages of tweets present on the dataset. English is the most prominent language in the dataset accounting for 67.59% of the total tweets. Figure 3 presents the number of tweets from each of the top 5 languages collected over time.

Table 3. Distribution of tweets, by language.

Language	English	Spanish	Portuguese	Bahasa	French	Others
Number of Tweets	530,653,121	105,593,454	32,085,122	22,166,492	17,491,687	77,128,691
Percentage	67.59	13.45	4.09	2.82	2.23	9.82

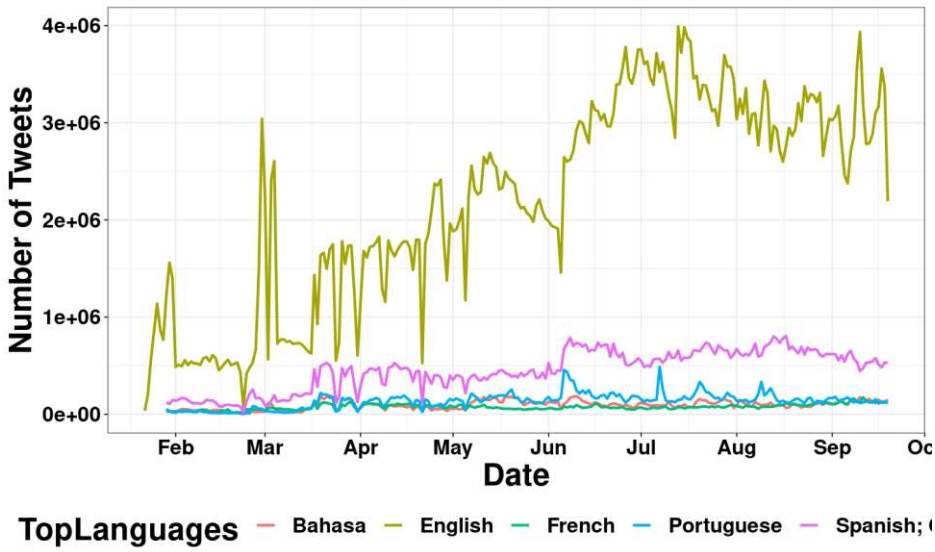

Figure 5. Tweet frequency across the top 5 observed languages.

We also collected information about the number of retweets, likes, and geolocation information. Table 4 shows the maximum and the median number of retweets and likes per month, as well as the total number of tweets with geolocation information. Because there is a large number of tweets that have zero likes, the median number of likes per month is always 0.

Table 4. Summary statistics for collected tweets**.**

Month	**Jan**	**Feb**	**Mar**	**Apr**	**May**	**Jun**	**Jul**	**Aug**	**Sept***
Geo	1,773	8,103	19,952	38,213	47,684	58,138	42,681	55,837	23,684
Max RT	674,151	469,739	1,064,693	649,823	1,007,616	790,652	608,890	2,183,434	1,925,489
MD RT	167	50	159	36	27	39	56	44	42
Max Like	334,802	637,589	1,255,858	662,005	929,811	882,693	1,287,117	860,162	839,689

**Until Sept. 19th, 2020*

While a total of 296,065 tweets with geolocation information are present on the dataset, this represents just 0.04% of the total number of tweets collected. Figure 6 presents the locations of the tweets with geolocation information since January 22nd, 2020.

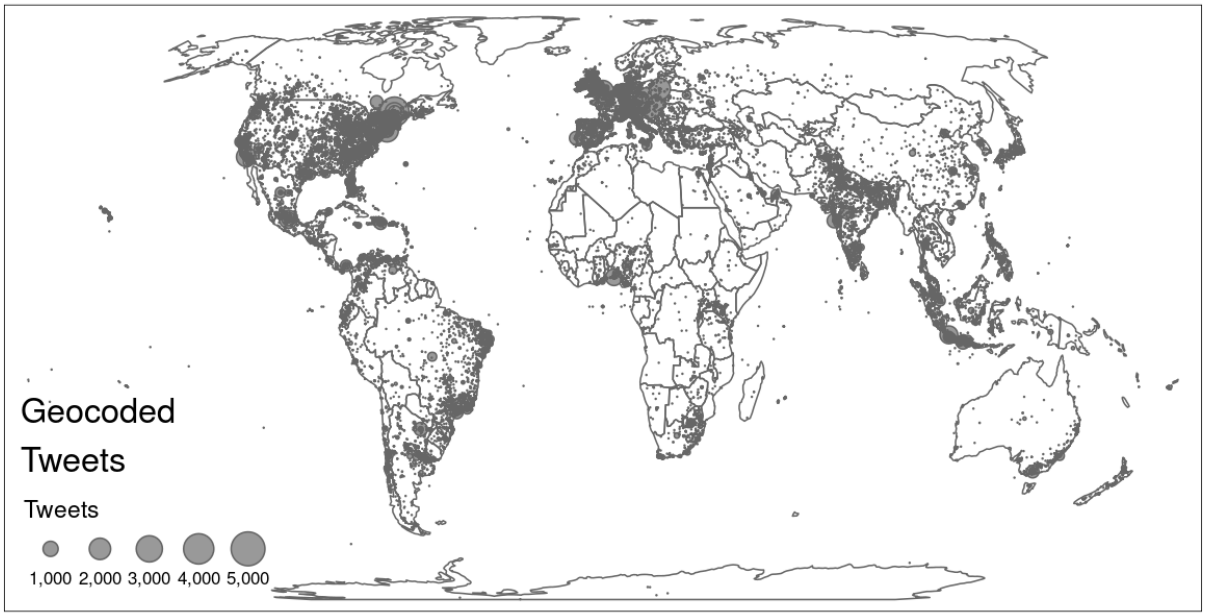

Figure 6. Map of tweets featuring geolocation information.

Similarly, the sentiment of all the English tweets was estimated using a state-or-the-art Twitter Sentiment algorithm. The dataset contains a total of 247,561,991 English tweets classified as negative (47.1 %), 40,575,026 as positive (7.8%) , and 194,179,806 as neutral (45.1%). Figure 7 shows the sentiment of all English-language tweets as of Sept 19th, 2020. Lastly, a Named Entity Recognition algorithm was used to extract topics of conversation about person, locations, organization, and others. Table 6 shows the top 5 mentions, hashtags, and named entities across the dataset.

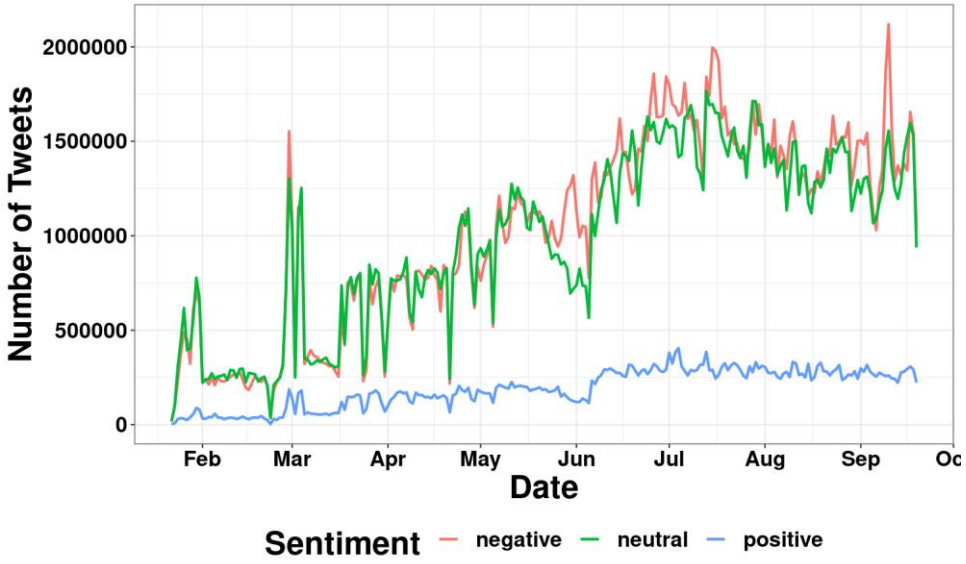

Figure 7. Sentiment of English-language tweets.

Table 5. Top 5 Mentions, hashtags, and named entities.

	1st	2nd	3rd	4th	5th
Mentions	@realDonaldTrump	@JoeBiden	@narendramodi	@DrRPNishank	@CNN
Hashtag	#covid19	#coronavirus	#covid	#covid-19	#stayhome
NER Person*	trump	god	donald trump	fauci	mike pence
NER Location*	china	us	wuhan	italy	india
NER* Organization	cdc	trump	nhs	world health organization	cnn
NER* Miscellaneous	coronavirus	chinese	covid-19	americans	democrats

* *The NER summary statistics only include tweets from 01/22/2020 to 04/20/2020*

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

4. Conclusion and Summary

Our main objective is to introduce and share with the research community a dataset of tweets related to the COVID-19 pandemic. We are continuously collecting and routinely updating the dataset with sentiment and NER annotations and producing summary files suitable for semantic network 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.

5. References

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

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/

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 (2019a) 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 (2019b) 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

Arora A, Bansal S, Kandpal C, Aswani R, Dwivedi Y (2019) Measuring social media influencer index-insights from Facebook, Twitter and Instagram. *Journal of Retailing and Consumer Services*, (49), 86–101.

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

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

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 in 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 (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. *The American Journal of Medicine*, 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/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

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. *The Journal of Bone and Joint Surgery*, 102(13): e64. DOI: 10.2106/JBJS.20.00610

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 (2020) ArCOV-19: The first Arabic COVID-19 Twitter dataset with propagation networks. https://arxiv.org/abs/2004.05861

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. *The Sociological Quarterly*, 59(1): 62-84.

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/

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

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).

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. *International Sociology*, 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

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. *Applied Intelligence*. https://doi.org/10.1007/s10489-020-01862-6

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

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

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

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

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 & Society*, *22*(4), 659-682.

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

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

Appendix 1. Openly available COVID-19 Twitter datasets

Citation	Approximate Tweets	Dates	Tweet ID	Time	Location	Sentiment	Topic	Other attributes
Abdul-Mageed, et al., 2020	1 X 10^9	2007 - May 15 2020	✔	✔				
Our dataset	785 X 10^6	22 Jan - Ongoing	✔	✔	✔	✔	✔	NER; Mentions; Hashtags
Banda, et al., 2020	728 X 10^6	27 Jan - Ongoing	✔	✔				Hashtag/mention summaries; 1,000 frequent terms
Chen, et al., 2020	623 X 10^6	28 Jan - Ongoing	✔					
Yang, Q., et al., 2020	105 X 10^6	01 Mar - 15 May 2020	✔			✔		
Gupta, et al., 2020	63 X 10^6	28 Jan - 01 July 2020	✔	✔	✔	✔	✔	User Metadata; Hashtags; Retweets
Ziems, et al., 2020	31 X 10^6	15 Jan - Ongoing	✔	✔	✔		✔	Sampled egonetworks
Gao, et al., 2020	25 X 10^6	20 Jan - 24 Mar 2020	✔	✔				
Dimitrov, et al., 2020	8.2 X 10^6	Oct 2019 - Apr 2020	✔	✔		✔		NER; Mentions; Hashtags; User Metadata; URLs
Alqurashi, et al., 2020	4.5 X 10^6	01 Jan – 30 Apr 2020	✔					
de Melo & Figueiredo, 2020	3.9 X 10^6	Jan - May 2020	✔	✔				Retweets; hashtags
Haouari, et al., 2020	1 X 10^6	27 Jan - 31 Mar 2020	✔					Propagation network of top 1,000 tweets by day
Feng & Zhou, 2020	650 X 10^3	25 Jan - 10 May 2020	✔	✔	✔	✔	✔	
Sarker, et al., 2020	472 X 10^3	01 Feb - Apr 2020						Self-reported COVID-19 symptoms
Cui & Lee, 2020	294 X 10^3	01 Dec - 01 Jul 2020	✔					User ID; Reply ID; Misinformation detection
Elhadad, et al., 2021	110 X 10^3	04 Feb - 10 Mar 2020	✔					Fact- checking annotation
Dharawat, et al., 2020	61 X 10^3	-						Health risk severity
Vidgen, et al., 2020	40 X 10^3	01 Jan - 10 Mar 2020					✔	
Mutlu, et al., 2020	14 X 10^3	04 Apr - 30 Apr 2020	✔					Human-coded stances on Hydroxychloroquine
Memon & Carley, 2020	4.5 X 10^3	29 Mar; 15/24 Jun 2020	✔	✔			✔	Tweets for users collected in this period

Figures

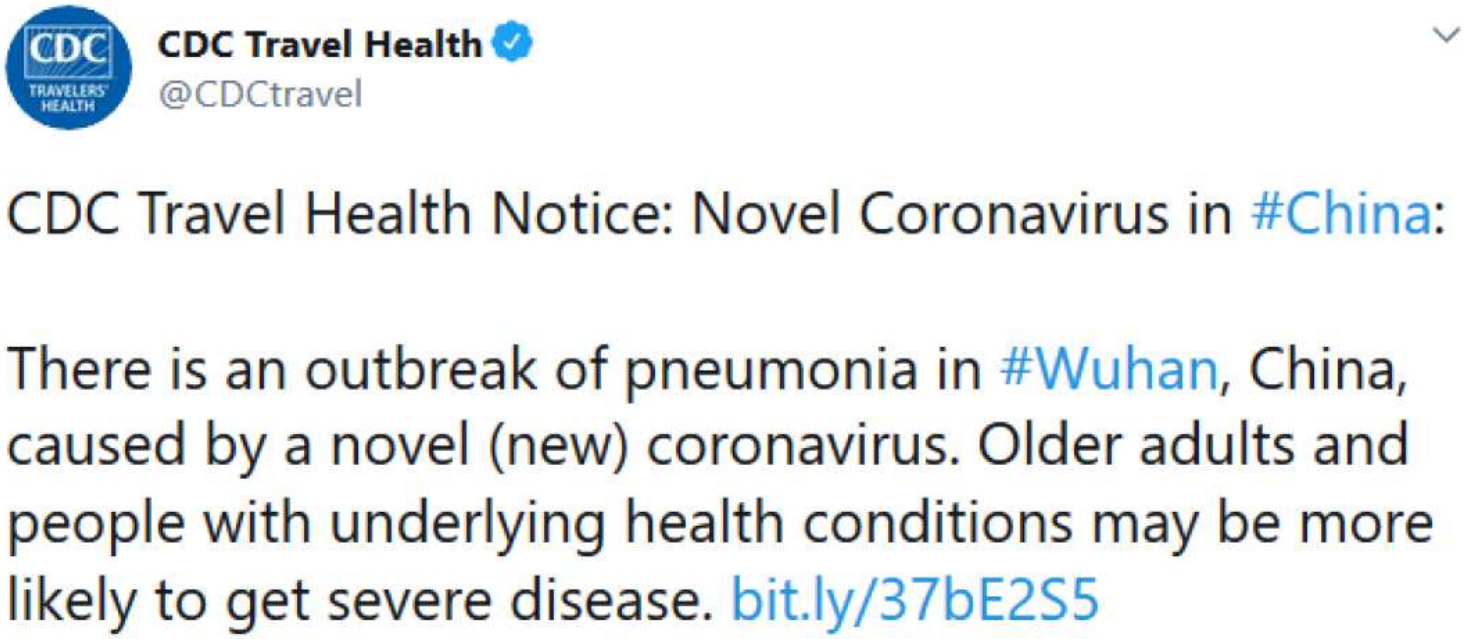

Figure 1

Example of Tweet Related to COVID-19

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

Figure 2

Example of dataset tables

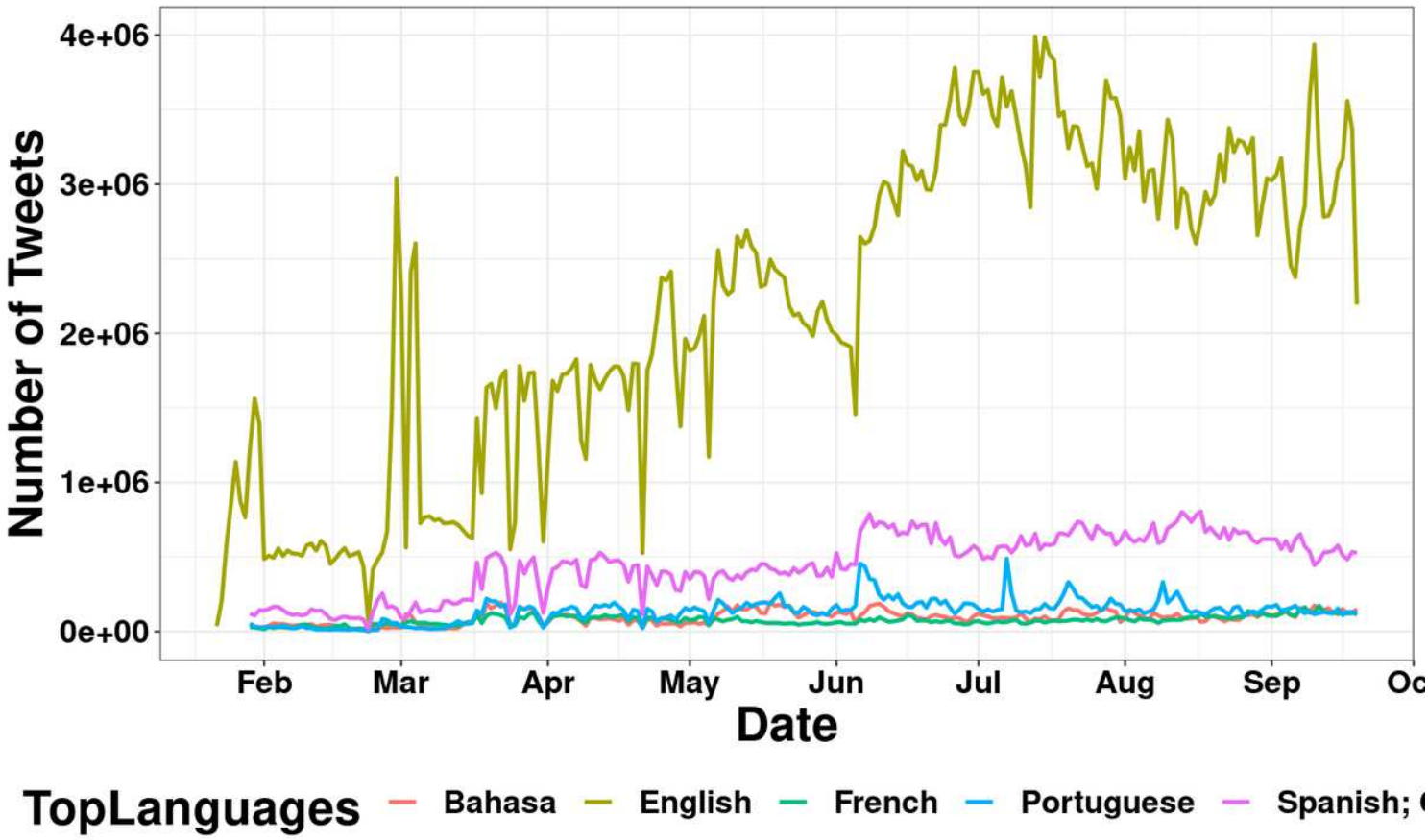

Figure 3

Tweets of top 5 languages

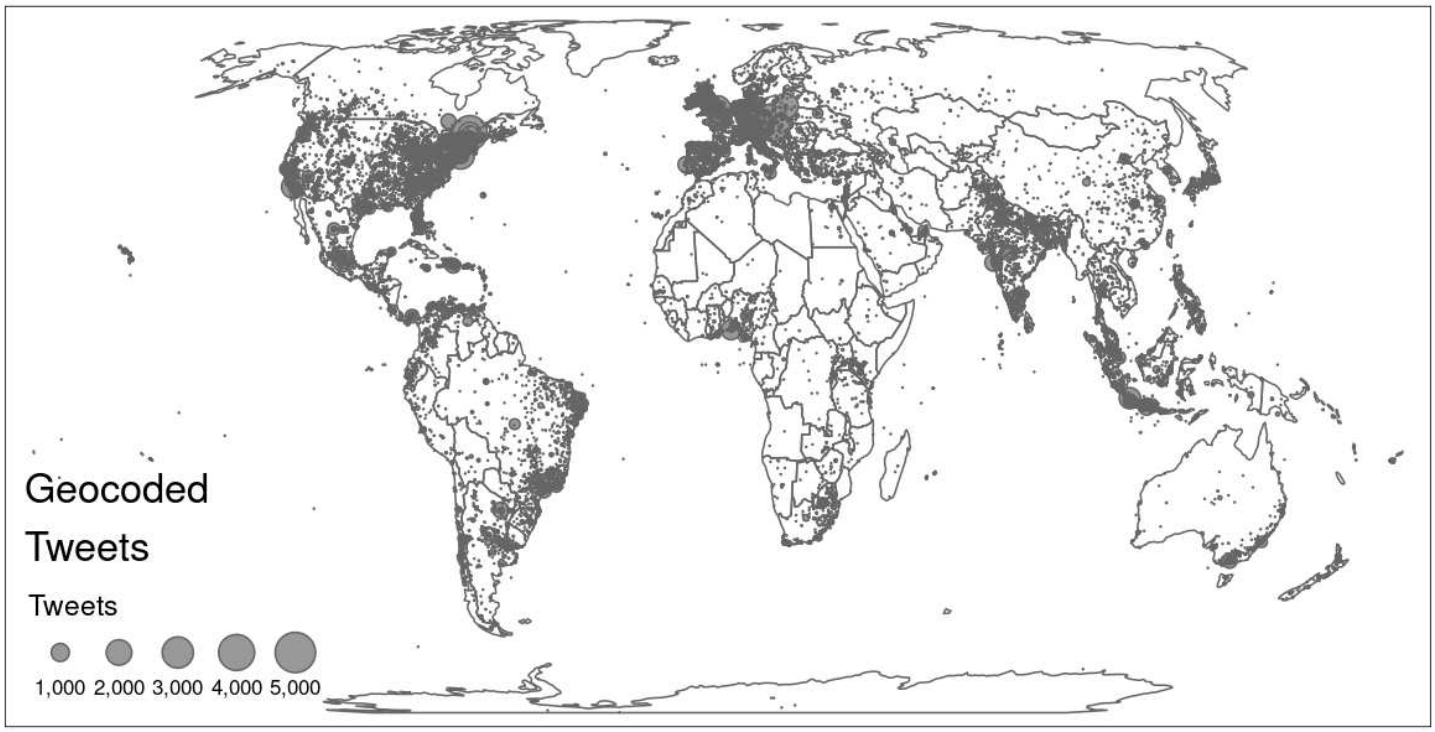

Figure 4

Map of Tweets with Geolocation. Note: The designations employed and the presentation of the material on this map do not imply the expression of any opinion whatsoever on the part of Research Square concerning the legal status of any country, territory, city or area or of its authorities, or concerning the delimitation of its frontiers or boundaries. This map has been provided by the authors.

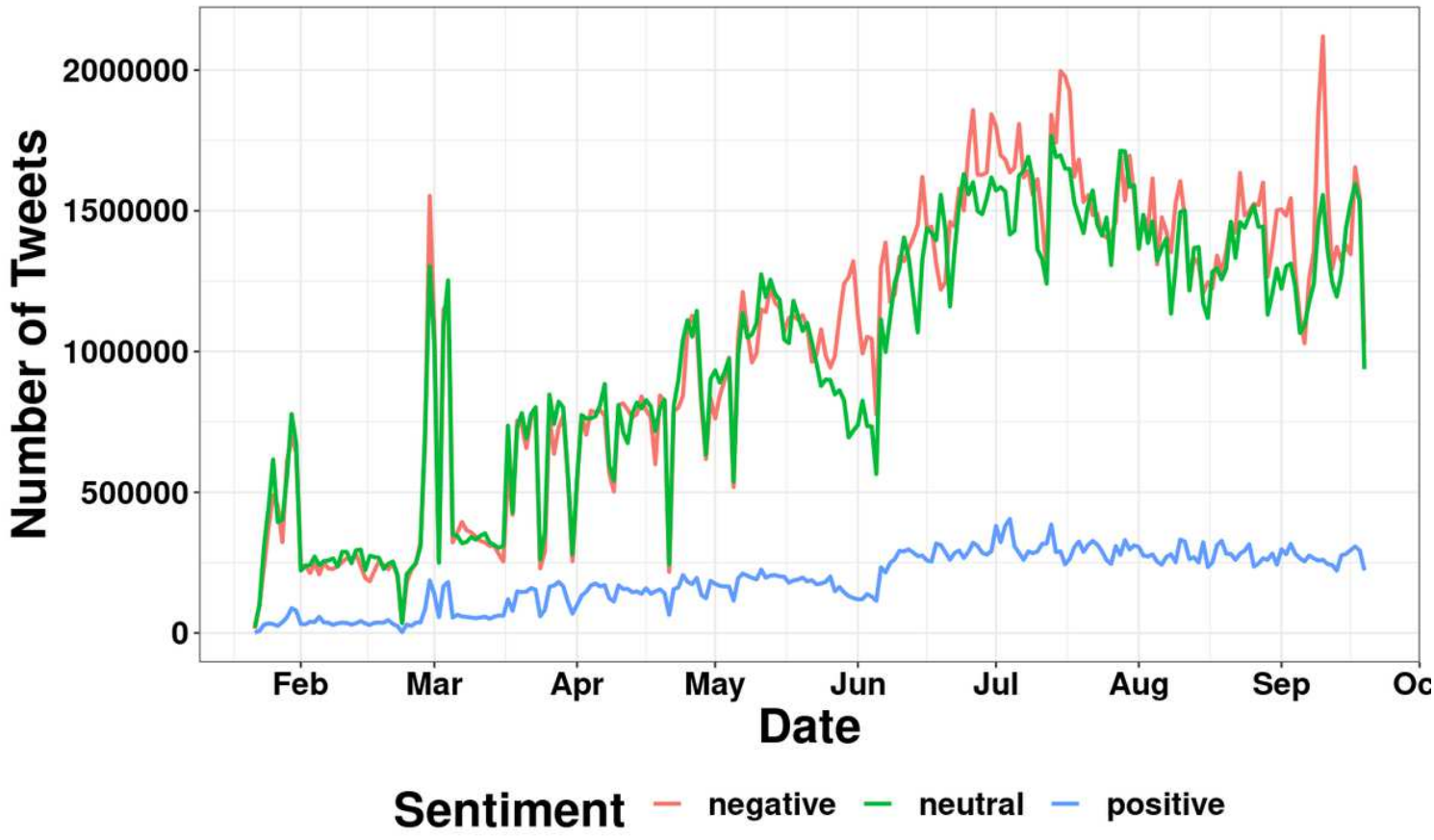

Figure 5

Sentiment of English-language tweets.

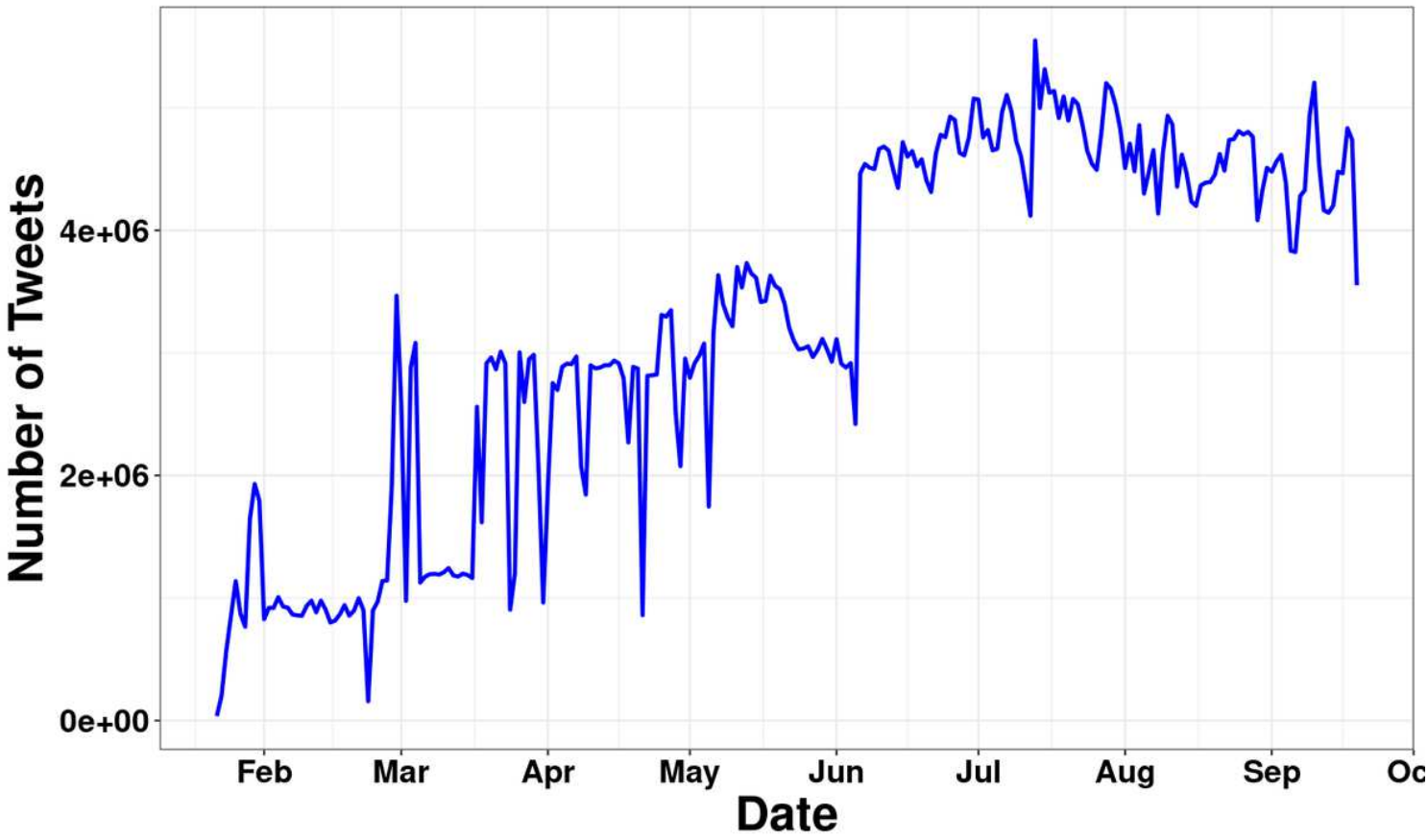

Figure 6

Tweets collected by day

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- lopezbecCOVID19TweetsDatasetCOVID19TweetsDataset.URL