# #COVID-19: The First Public Coronavirus Twitter Dataset

**Emily Chen, Kristina Lerman, Emilio Ferrara**
University of Southern California, Information Sciences Institute
4676 Admiralty Way, #1001
Marina del Rey, CA 90292
Corresponding author: emiliofe@usc.edu

## Abstract

At the time of this writing, the novel coronavirus (COVID-19) pandemic outbreak has already put tremendous strain on many countries' citizens, resources and economies around the world. Social distancing measures, travel bans, self-quarantines, and business closures are changing the very fabric of societies worldwide. With people forced out of public spaces, much conversation about these phenomena now occurs online, e.g., on social media platforms like Twitter. In this paper, we describe a multilingual coronavirus (COVID-19) Twitter dataset that we have been continuously collecting since January 22, 2020. We are making our dataset available to the research community: `https://github.com/echen102/COVID-19-TweetIDs` It is our hope that our contribution will enable the study of online conversation dynamics in the context of a planetary-scale epidemic outbreak of unprecedented proportions and implications. This dataset could also help track scientific coronavirus misinformation and unverified rumors, or enable the understanding of fear and panic — and undoubtedly more. Ultimately, this dataset may contribute towards enabling informed solutions and prescribing targeted policy interventions to fight this global crisis.

## Introduction

The first cases of a novel coronavirus disease (officially named COVID-19 by the World Health Organization on February 11, 2020) were reported in Wuhan, China in late December 2019; the first fatalities were reported in early 2020. The fast rising infection rates and death toll led the Chinese government to quarantine the city of Wuhan on January 23, 2020.[1] During this period, other countries began reporting their first confirmed cases of the disease, and on January 30, 2020 the World Health Organization (WHO) announced a Public Health Emergency of International Concern. With more countries reporting cases of the disease, and infections rapidly escalating in some regions of the world, including South Korea, Iran and Italy, WHO has upgraded COVID-19 to a pandemic.[2] As of the time of this writing, COVID-19 has been reported in 146 countries, leaving governments all over the world scrambling for ways to contain the disease and lessen its adverse consequences to their people's health and economy.[3]

Preventative measures implemented by national, state and local governments now affect the daily routines of millions of people worldwide. "Social distancing," the most widely used of such measures, aims to curtail new infections by reducing physical contact between people. Social distancing measures has led to the cancellation of sporting events and conferences, closures of schools and colleges, and has forced many businesses to require their employees to work from home. As more and more social interactions move online, the conversation around COVID-19 has continued to expand, with growing numbers turning to social media for both information and company. Platforms such as Twitter have become central to the technological and social infrastructure that allows us to stay connected even during crises.

We describe a dataset about COVID-19-related online conversations that we are sharing with the research community. We hope that this data will spur new research about the social dimensions of the pandemic. We began collecting data in real-time from Twitter, with the earliest tweets from January 22, 2020, tracking COVID-19 related keywords and accounts. Here we describe the data collection methods, document initial data statistics, and provide information about how to obtain and use the data.

## Data Collection

We began actively collecting tweets from January 28, 2020, leveraging Twitter's streaming API and Tweepy[4] to follow specific keywords and accounts that were trending at the time. When we started collecting tweets, we also used Twitter's search API on the same keywords to gather related historical tweets. Thus the earliest tweets in our collection date back to January 22, 2020. Since then, we have incrementally added keywords and accounts to follow based on the conversations occurring on Twitter at any time. We have collected over 50 million tweets from the inception until March 16, 2020, constituting roughly 450 GB of raw data.

Our data collection will continue uninterruptedly for the foreseeable future. Future versions of this paper will document and keep record of this data collection process. As the

[1] https://www.nytimes.com/article/coronavirus-timeline.html
[2] https://www.who.int/emergencies/diseases/novel-coronavirus-2019/events-as-they-happen
[3] https://coronavirus.jhu.edu/map.html
[4] https://www.tweepy.org/

| Keyword | Tracked Since |
|---|---|
| Coronavirus | 1/22/2020 |
| Koronavirus | 1/22/2020 |
| Corona | 1/22/2020 |
| CDC | 1/22/2020 |
| Wuhancoronavirus | 1/22/2020 |
| Wuhanlockdown | 1/22/2020 |
| Ncov | 1/22/2020 |
| Wuhan | 1/22/2020 |
| N95 | 1/22/2020 |
| Kungflu | 1/22/2020 |
| Epidemic | 1/22/2020 |
| Outbreak | 1/22/2020 |
| Sinophobia | 1/22/2020 |
| China | 1/22/2020 |
| Covid-19 | 2/16/2020 |
| Corona virus | 3/2/2020 |
| Covid | 3/6/2020 |
| Covid19 | 3/6/2020 |
| Sars-cov-2 | 3/6/2020 |
| COVID–19 | 3/8/2020 |
| COVD | 3/12/2020 |
| Pandemic | 3/12/2020 |

Table 1: Keywords that we are actively tracking in our Twitter collection (v1.0 — March 16, 2020).

pandemic continues to run its course, we anticipate that the amount of data will grow significantly.

In compliance with Twitter's Terms & Conditions, we are unable to publicly release the text of the collected tweets. We are, therefore, releasing the tweet IDs that researchers can then use to retrieve the full tweet object. There are several easy to use tools that have been developed for such purposes, including the *Hydrator*.[5]

### Tracked Keywords and Accounts

By continuously monitoring Twitter's trending topics, keywords and sources associated with COVID-19, we did our best to capture conversations related to the coronavirus outbreak. Twitter's streaming API returns any tweet containing the keyword(s) in the text of the tweet, as well as in its metadata; so it is not always necessary to have each permutation of a specific keyword in the tracking list. For example, the keyword "Covid" will return tweets that contain both "Covid19" and "Covid-19". We list the keywords and accounts that we are following in Table 1 and Table 2 respectively, along with the date we began tracking them. Due to the evolving nature of the pandemic and online conversations, these tables will expand as we continue to monitor Twitter for additional keywords and accounts to add to our tracking list.

[5] https://github.com/DocNow/hydrator

| Account Name | Tracked Since |
|---|---|
| PneumoniaWuhan | 1/22/2020 |
| CoronaVirusInfo | 1/22/2020 |
| V2019N | 1/22/2020 |
| CDCemergency | 1/22/2020 |
| CDCgov | 1/22/2020 |
| WHO | 1/22/2020 |
| HHSGov | 1/22/2020 |
| NIAIDNews | 1/22/2020 |

Table 2: Account names that we are actively tracking in our Twitter collection (v1.0 — March 16, 2020).

| Language | ISO | No. tweets | % total |
|---|---|---|---|
| English | en | 5,508,304 | 61.76% |
| Spanish | es | 1,167,172 | 13.09% |
| French | fr | 388,481 | 4.36% |
| Thai | th | 352,902 | 3.96% |
| Italian | it | 219,572 | 2.46% |
| (undefined) | und | 208,908 | 2.34% |
| Indonesian | in | 201,821 | 2.26% |
| Portuguese | pt | 169,599 | 1.9% |
| Japanese | ja | 145,985 | 1.64% |
| Turkish | tr | 134,173 | 1.5% |

Table 3: Breakdown of the most popular languages and the amount of associated tweets (v1.0 — March 16, 2020).

## Data & Access Modalities

### Release v1.0 (March 16, 2020)

We provide some preliminary statistics of the first dataset release. This initial dataset includes tweets collected from March 5, 2020 through 5 pm UTC March 12, 2020, **8,919,411** tweets in all. The language breakdown of the tweets can be found in Table 3. The keywords and accounts that were followed during this time-frame can be identified by referencing Tables 1 and 2. Some of the keywords may appear earlier than the initial listed track date in Table 1, as we systematically ran the same keywords through Twitter's search API to collect past instances of the keywords shortly after adding the keywords to be tracked in real-time.

### Access

The dataset is available on Github at this address:
**https://github.com/echen102/COVID-19-TweetIDs**
The dataset is released in compliance with the Twitter's Terms & Conditions.[6] This dataset is still being continuously collected and routinely updated. To request access, interested researchers will need to agree upon the terms of usage dictated by the chosen license.

If you have technical questions about the data collection, please contact Emily Chen at echen920@usc.edu
If you have any further questions about this dataset please contact Dr. Emilio Ferrara at emiliofe@usc.edu

**Acknowledgments**. The authors gratefully acknowledge support from DARPA (contract no. W911NF-17-C-0094).

[6] https://developer.twitter.com/en/developer-terms/ag