Data Descriptor

A Geo-Tagged COVID-19 Twitter Dataset for 10 North American Metropolitan Areas over a 255-Day Period

Sara Melotte [1,†] and Mayank Kejriwal [1,‡]

[1] University of Southern California, Information Sciences Institute
* Correspondence: kejriwal@isi.edu
† melotte@usc.edu
‡ kejriwal@isi.edu

Abstract: One of the unfortunate findings from the ongoing COVID-19 crisis is the disproportionate impact the crisis has had on people and communities who were already socioeconomically disadvantaged. It has, however, been difficult to study this issue at scale and in greater detail using social media platforms like Twitter. Several COVID-19 Twitter datasets have been released, but they have very broad scope, both topically and geographically. In this paper, we present a more controlled and compact dataset that can be used to answer a range of potential research questions (especially pertaining to computational social science) without requiring extensive preprocessing or tweet-hydration from the earlier datasets. The proposed dataset comprises tens of thousands of geotagged (and in many cases, reverse-geocoded) tweets originally collected over a 255-day period in 2020 over 10 metropolitan areas in North America. Since there are socioeconomic disparities within these cities (sometimes to an extreme extent, as witnessed in 'inner city neighborhoods' in some of these cities), the dataset can be used to assess such socioeconomic disparities from a social media lens, in addition to comparing and contrasting behavior across cities.

Keywords: COVID-19; Twitter; Geo-Tagged; Metropolitan; Computational Social Science

1. Summary

In addition to its medical consequences, the ongoing COVID-19 crisis has also revealed (if not exacerbated) deep inequalities in our society [1], [2], [3], [4]. Early surveys and results show that people and communities of lower socioeconomic status have been disproportionately affected. However, it has been difficult to study this issue in greater detail using social media sources like Twitter. While several COVID-19 Twitter datasets have been released [5], [6], [7], [8], they are broad-ranging (both topically and geographically). What is missing is a carefully controlled dataset that would enable computational social scientists in specific contexts to study the issue from a social media lens without much hassle. At the same time, rather than reinvent the wheel and collect raw data from scratch, it should be possible to use some of these earlier larger datasets as a *starting point* for constructing the more controlled (and also, appropriately augmented) dataset.

In this paper, we address these desiderata by presenting a dataset that, while compact, contains many tens of thousands of tweets that comprise a sub-set of the broader GeoCOV19Tweets dataset, originally obtained by filtering English tweets from the Twitter streaming API by using a continuously updated, expansive list of keywords and hashtags [7]. As of this writing, the GeoCOV19Tweets Twitter feed is monitored using 90+ keywords and hashtags commonly used when referencing the pandemic. Although only English tweets were gathered, the data collection has global span. Each collection starts between 10:00-11:00hrs GMT+5:45 every day [9]. The data collection started on March 20, 2020 and has been updated daily with newly collected tweet IDs.

Unlike the GeoCOV19Tweets dataset, our dataset has further filtered the tweets based on location of origin in one of the 10 most populous cities in the United States

Figure 1. An example JSON dictionary fragment representing a tweet (originating in Los Angeles) in our dataset, with the metadata.

and Canada[1]. Tweets in our dataset are accompanied by the tweet's date, hashtags, and the city, state, and place type where the tweet originated. We publish a separate file for each metropolitan area, written out in a compact key-value file format. We retain the sentiment scores included with the original GeoCOV19Tweets dataset, and we also augment the dataset with extracted hashtags, since there are a number of computational social science studies that primarily rely on hashtags (and can hence avoid hydrating the tweets, if obtaining the original tweet text is not necessary). Our primary goal in publishing this dataset is to enable social scientists and digital humanities scholars with a less technical background to study COVID-19 in metropolitan contexts, over a longitudinal period, through a social media lens. For this reason, our dataset is compact and places a high premium on accurate geotagging, the details of which are described subsequently.

2. Data Description

The DOI of the dataset is *10.5281/zenodo.4434972* and it is available publicly, with documentation, at https://zenodo.org/record/4434972#.YKA7bJNKhBw (accessed on May 15, 2021). It is licensed under *Creative Commons Attribution 4.0 International*.

The data release comprises 10 Java Script Object Notation (JSON) files, one for each of the 10 metropolitan areas. The upper-level JSON object within the file is a list, each element of which is a dictionary (also technically a JSON, since JSON is defined recursively). Each dictionary represents a tweet. In addition to containing geographic metadata, the tweet also retains the sentiment score in the GeoCOV19Tweets[2] data.

For each tweet, the following (mnemonically named) fields are recorded: tweet ID, sentiment score, date, hashtags, city, state, and place_type[3]. A fragment is shown in Figure 1.

Concerning quality, we note that since the dataset is an augmented subset of GeoCOV19Tweets, the coverage of the dataset is bound above by the coverage of GeoCOV19Tweets. However, one reason why we used GeoCOV19Tweets as the original source is that it seems to have excellent coverage due to its expansive use of COVID-19 related hashtags [7]. Beyond coverage, we discuss other qualitative properties of the dataset in *Statistics and usage notes*. As with all Twitter data, certain well-known caveats always apply when using such data, including influence of bots and disinformation [10].

[1] Montreal is excluded due to its significant French-speaking population, as discussed in *Collection Methodology*.

[2] According to its documentation, GeoCOV19Tweets itself used the Python-based TextBlob package (on the text of the tweet) to automatically obtain a sentiment score.

[3] While the place_type usually contains a string indicating the type of the place (e.g., 'city'), in some cases, it may store the zipcode due to the need for a reverse geocoding service, as discussed in *Collection Methodology*.

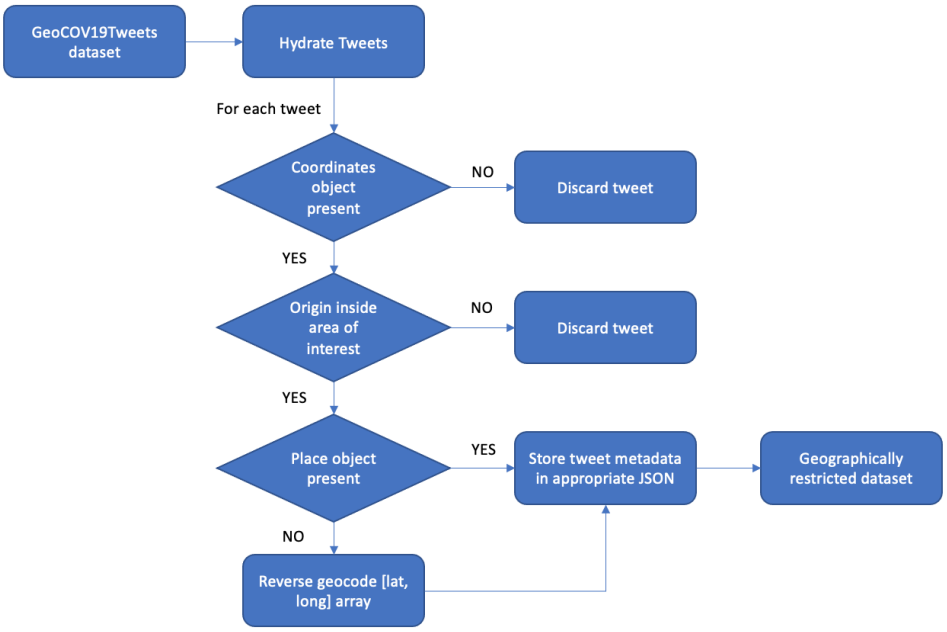

Figure 2. A workflow illustrating the methodology behind data processing and collection as applied to the underlying GeoCOV19Tweets dataset to obtain the proposed dataset.

3. Methods

3.1. Preliminaries: GeoCOV19Tweets Dataset

The GeoCOV19Tweets dataset [7] is used as the super-set of tweets that we hydrate, filter and augment to obtain our dataset. Within GeoCOV19Tweets, each day is represented by a separate Comma Separated Values (CSV) file, with each record within the file containing the tweet ID and automatically generated sentiment score. Under Twitter's terms and conditions, only by hydrating the tweet IDs can additional information about the tweets be obtained. The days of October 27, 2020 to October 28, 2020 do not include sentiment analysis (only tweet IDs) and are therefore omitted in the creation of the presented dataset as well.

The GeoCOV19Tweets dataset includes only tweets that are geo-tagged, allowing for geographical analysis of the tweets' metadata. When a Twitter user grants access to their location via Global Positioning System (GPS), the geo-coordinate data is added to the tweet location, giving rise to various geo-objects [9]. Some of these objects are used to further filter the data and create the presented dataset. We provide more details subsequently in *Collection Methodology*.

The sentiment scores are generated using the TextBlob's Sentiment Analysis tool[4]. The score is a continuous value in a range of [-1, 1], where more positive (negative) values signify positive (negative) sentiment and 0 signifies a Neutral sentiment [11]. Prior to computing sentiment scores, the tweets are preprocessed by cleaning the hash symbol (#), mention symbol (@), URLs, extra spaces, and paragraph breaks. Punctuation, emojis, and numbers are included. No other preprocessing is done.

3.2. Collection Methodology

Figure 2 summarizes the workflow of our collection methodology. Details on individual important steps are described below.

[4] https://textblob.readthedocs.io/en/dev/

3.2.1. Hydrating Tweets

As a first step, we hydrate tweet IDs in GeoCOV19Tweets from the March 20, 2020 through the December 1, 2020 period (255 days) using the Python *twarc* library[5], followed by further filtering based on the location of origin. From the Twitter Developer API, the tweet ID, UTC date and time of creation, hashtags, coordinates, city, state, and place type are collected. The period October 27, 2020–October 28, 2020 is skipped because no sentiment scores were provided in the underlying GeoCOV19Tweets dataset, as noted earlier. The month of March includes data for an 11-day period only[6], whereas all other months in this dataset comprise the whole month.

3.2.2. Determining Tweet Origin

One of the issues with the tweets hydrated from GeoCOV19Tweets is that some tweets contain a user-defined location tag that may be different from the origin of the tweet. We are interested in the latter rather than the former. To enforce this constraint, we only consider tweets having a populated (i.e., non-null) "coordinates" object in the metadata. This precludes inclusion of tweets that may be assigned a user-defined location in the "place" object associated with the tweet, even though the tweet *itself* did not originate from that place. However, tweets returned from the Twitter Developer API having the "coordinates" object defined, populate the "place" object corresponding to the location indicated by the "coordinates" object. Therefore, an important processing step upon hydration is to filter out tweets that do not have a "coordinates" object defined in the metadata.

3.2.3. Reverse-Geocoding

As mentioned earlier, although the Twitter Developer API deduces the "place" object from the "coordinates" object associated with the tweet, sometimes, the "place" object is *None* in a tweet even though the "coordinates" object is defined. In these instances, we reverse-geocode the latitude and longitude in the "coordinates" object using the *Geocodio* tool[7].

Geocodio's API allows both forward- and reverse-geocoding within the United States and Canada, returning up to five possible matches ranked by an accuracy score between 0.00 and 1.00. When the geocoding service is needed (i.e. when a "coordinates" object, but not a "place" object exists within the tweet metadata), we use the reverse-geocoded result with the highest accuracy score to deduce the location's city, state, zip code, and country.

Unlike the Twitter Developer API, Geocodio is unable to provide the "place_type" of the latitude-longitude location. The "place_type" field returned by the Twitter API contains a description of the tweet's origin (e.g., "city" or "admin"). However, Geocodio can return the location's zipcode. For this reason, the "place_type" field in the presented dataset may either contain the "place_type" extracted directly from the Twitter API (which is a string) or the zipcode obtained from Geocodio. We note, however, that in most cases, the "coordinates" object did not need to be reverse-geocoded and we were simply able to use the populated "place" object provided directly in the tweet.

3.2.4. Selecting Metropolitan Areas

The 10 most populous cities in the United States and Canada are chosen as the areas of interest (Figure 3). Although Montreal is the sixth most populous city in the United States-Canada region, it is disregarded due to its significant French-speaking population[8]

[5] https://scholarslab.github.io/learn-twarc/

[6] The GeoCOV19Tweets dataset begins on March 20, 2020.

[7] https://www.geocod.io/

[8] According to https://www12.statcan.gc.ca/census-recensement/2016/dp-pd/prof/details/page.cfm?Lang=E&Geo1=CMACA&Code1=462&Geo2=PR&Code2=01&Data=Count&SearchText=montreal&SearchType=Begins&SearchPR=01&B1=All&TABID=1

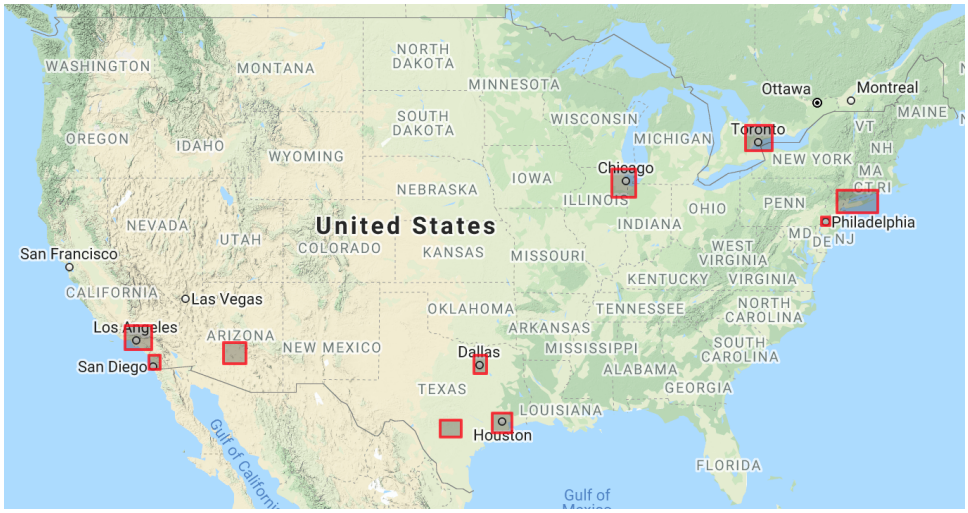

Figure 3. Bounding rectangles for New York, Los Angeles, Toronto, Chicago, Houston, Phoenix, Philadelphia, San Antonio, San Diego, and Dallas.

	Top Left	Bottom Right	Tweet Count	Percentage (%)
New York	(41.415634, -74.485085)	(40.411124, -71.853181)	20979	40.3163
Los Angeles	(34.820691, -118.946542)	(33.602688, -117.275379)	13893	26.6988
Toronto	(44.383080, -80.114152)	(43.284905, -78.473654)	5505	10.5792
Chicago	(42.391280, -88.501901)	(41.122449, -87.009653)	3171	6.0939
Houston	(30.218201, -95.934175)	(29.136616, -94.729970)	2220	4.2663
Phoenix	(33.942208, -112.752517)	(32.812873, -111.362060)	1123	2.1581
Philadelphia	(40.158714, -75.403683)	(39.777982, -74.913390)	1413	2.7154
San Antonio	(29.850468, -99.185990)	(28.902995, -97.884110)	697	1.3395
San Diego	(33.249462, -117.432605)	(32.533032, -116.733257)	1411	2.7116
Dallas	(33.249352, -97.130478)	(32.326729, -96.342209)	1624	3.1209
		Total	**52036**	

Table 1: Coordinates (lat, long) of bounding rectangles for each selected metropolitan area, along with tweet counts and percentages.

and the retention of exclusively English tweets in the original GeoCOV19Tweets dataset. The presented dataset is intended to best represent the city's Twitter-users' sentiments toward COVID-19 based on English tweets only. The 10 metropolitan areas, in order of decreasing population size[9], selected for this dataset are: New York, Los Angeles, Toronto, Chicago, Houston, Phoenix, Philadelphia, San Antonio, San Diego, and Dallas.

3.2.5. Location-Based Filtering

To determine whether a tweet originates from one of the 10 areas of interest (Figure 3), bounding rectangles are drawn around the city and its surrounding neighborhoods. Any tweet that originates within or on the bounding box is labeled with the respective metropolitan area. The selected coordinates for each metropolitan area are tabulated in Table 1. As previously discussed, only tweets within GeoCOV19Tweets with a known and exact location of origin in one of these metropolitan areas (as opposed to user-defined tag) are retained in the metropolitan-specific JSON files. Tweet counts and percentages collected for each metropolitan area are tabulated in Table 1.

3.3. Related Datasets

A number of datasets related to COVID-19 have been released in the last year. We note the ones that are particularly related to the presented dataset below, and also

[9] According to https://www.census.gov/newsroom/press-releases/2020/south-west-fastest-growing.html and https://www12.statcan.gc.ca/census-recensement/2016/as-sa/98-200-x/2016001/98-200-x2016001-eng.cfm.

describe the value that our dataset provides that helps complement the value of these other datasets.

The *USC Dataset* by [6] tracks social media discourse about the COVID-19 pandemic. While broad (comprising more than a hundred million tweets at present), the dataset does not limit itself to geotagged tweets, and it is difficult to obtain this data without first hydrating the tweets. The dataset is therefore useful for large scale studies tracking COVID discourse on social media (as the authors present its primary use case to be) but not as much for more constrained studies within a specific locational and topical context. Other similar examples include the COVID-19 Twitter Dataset [5] and TweetsCOV19 [8]. Earlier, we also provided details on GeoCOV19Tweets [7], which served as the primary super-set on which the presented dataset is based.

Other examples that are designed for studying specific topics or are in specific languages include work by [12], [13], [14], and [15]. However, to our knowledge, there is no dataset with the goal of studying longitudinal and cross-sectional differences and similarities between metropolitan samples in the COVID-19 context.

3.4. Ethical Considerations

We have not identified any ethical concerns with the dataset, since the tweets were scraped from the public API, and are themselves a small subset of another publicly available data source (GeoCOV19Tweets). We have not released user account information or the text of the tweet, in keeping with Twitter's terms and conditions.

3.5. Possible Compliance with FAIR

The presented dataset, with its metadata, is *findable* as it has been hosted on Zenodo and assigned a globally unique and eternally persistent DOI. It has also been described with rich metadata, and indexed in a searchable resource. The metadata also specify the data identifier.

The dataset is also *accessible* and *interoperable*, the latter due to the use of a formal, accessible, shared and broadly applicable language (JSON) for knowledge representation. Finally, the data is *re-usable* as it is also associated with provenance, since it has been derived from GeoCOV19Tweets, and the tweets are also accompanied by their IDs. The dataset also has a plurality of accurate and relevant attributes (including location, hashtags and sentiment score) that could be used to answer a range of computational social science questions.

3.6. Statistical Summary

Out of the 272,404 tweets collected and hydrated from the GeoCOV19Tweets dataset between March 20, 2020 and December 1, 2020 (255-day period), 52,036 tweets remain in this dataset. This represents 19.103% of the original tweets collected. The highest percentage of tweets in the overall dataset comes from the New York metropolitan area, followed by Los Angeles and Toronto (Figure 3). This trend is correlated with their relative population sizes. For example, New York metropolitan area makes up 40.3163% of the dataset with a total of 20,979 tweets.

4. Detailed Statistics and Usage Notes

4.1. Statistics on Sentiment Scores

The average sentiment score for the overall dataset is 0.1279, indicating a small positive sentiment. The lowest average sentiment score is found in May and the highest in November. The months of July and September tie at 0.1403.

Throughout the 255-day period for which tweets across 10 metropolitan areas are collected, the average sentiment remained positive. Even the lowest average sentiment,

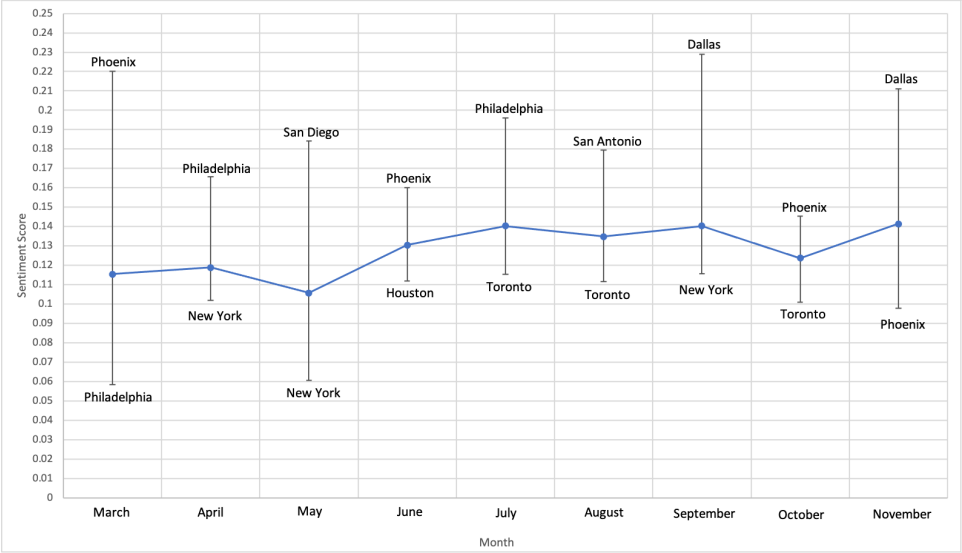

Figure 4. Sentiment scores versus month over all the tweets in our dataset. We also illustrate (for each month) the metropolitan areas with the lowest and highest average sentiment scores.

measured in March[10] in Philadelphia (0.0584), indicates a positive sentiment towards the COVID-19 pandemic on Twitter (Figure 4). The most positive average sentiment is recorded in Dallas in September (0.2288). Los Angeles and Chicago are the only metropolitan areas whose monthly average is neither the lowest nor highest average sentiment in any month throughout the period analyzed.

Obviously, these results should not be meant to imply that people were overall feeling positive about this crisis in North America, where the response to the crisis has been particularly criticized. It may mean that either the use of TextBlob should be revisited (for anyone looking to re-hydrate the tweets and do deeper textual analysis) or that sentiment scores across the tweets should be interpreted more relatively. It may also indicate biases that deserve further study, and that the provision of this dataset can help support, especially if the metropolitan area needs to be controlled for.

4.2. Statistics on Hashtags

In the overall dataset, 31,041 tweets included hashtags (59.6529%). The 10 most commonly included hashtags are illustrated in Figure 5. The prevalence of each hashtag is calculated with respect to the number of tweets in the overall dataset, while the lowest and highest prevalence of each hashtag are calculated with respect to the number of tweets in each of the 10 metropolitan areas, providing information about geographic trends in hashtag usage.

Unsurprisingly, the most commonly used hashtag is "covid19", with a prevalence of 11.4133% across the overall dataset. The hashtags "SaveTheWorld", "BillionShieldsChallenge", and "BillionShields" are prevalent in New York metropolitan area but do not appear in 7 of the 10 selected metropolitan areas. Similarly, "faceshield" is used by Twitter users in New York metropolitan area but not at all by users in Phoenix or San Antonio. This may be explained by the large number of tweets collected in New York metropolitan area.

4.3. Possible Use-Cases

We hypothesize that the dataset can be used to address a range of research questions:

[10] As mentioned, the original dataset began sampling tweets on March 20, 2020. The average sentiment score for March is therefore taken over an 11-day period.

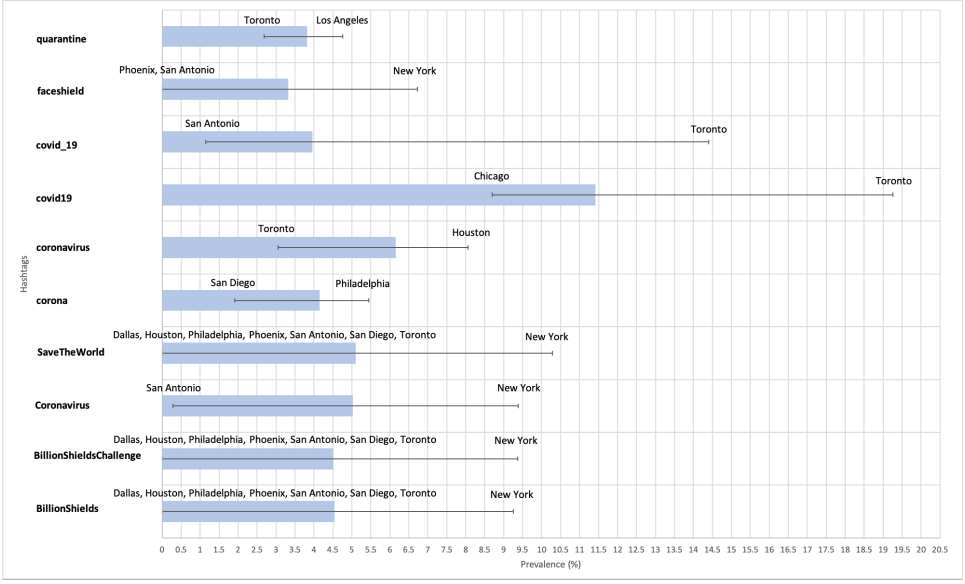

Figure 5. The 10 most prevalent hashtags (determined over all the tweets in our dataset). We also illustrate, for each hashtag, the metropolitan area with the lowest and highest prevalence in the corresponding metropolitan area.

1. Given that different metropolitan areas were impacted differently by COVID-19 (in particular, New York was hit hard in the early days), how is this impact reflected in social media?
2. Can tweets from areas (with different socioeconomic profiles) within metropolitan cities shed light on how socioeconomic status is correlated with COVID-19 impacts, and how such correlations manifest on social media? While limited surveys and studies have confirmed that COVID-19 disproportionately affected lower socioeconomic-status groups, to our knowledge, a full study through a social media lens has not yet emerged.
3. Given policy measures that were enacted in different cities over time, what can we say about longitudinal differences (especially in terms of sentiment) between these cities?

We note again that an important advantage of this dataset is that some of these questions can be answered relatively quickly, due to the far lower number of tweets that would have to be hydrated. For other questions, only sentiment analysis or hashtags may be necessary, which would require no hydration at all as we provide these metadata.

References

1. Bowleg, L. We're not all in this together: on COVID-19, intersectionality, and structural inequality, 2020.
2. Patel, J.; Nielsen, F.; Badiani, A.; Assi, S.; Unadkat, V.; Patel, B.; Ravindrane, R.; Wardle, H. Poverty, inequality and COVID-19: the forgotten vulnerable. *Public health* **2020**, *183*, 110.
3. Blundell, R.; Costa Dias, M.; Joyce, R.; Xu, X. COVID-19 and Inequalities. *Fiscal Studies* **2020**, *41*, 291–319.
4. Abedi, V.; Olulana, O.; Avula, V.; Chaudhary, D.; Khan, A.; Shahjouei, S.; Li, J.; Zand, R. Racial, economic, and health inequality and COVID-19 infection in the United States. *Journal of racial and ethnic health disparities* **2020**, pp. 1–11.
5. Gruzd, A.; Mai, P. COVID-19 Twitter Dataset, 2020. doi:10.5683/SP2/PXF2CU.
6. Chen, E.; Lerman, K.; Ferrara, E. Tracking Social Media Discourse About the COVID-19 Pandemic: Development of a Public Coronavirus Twitter Data Set. *JMIR Public Health Surveill* **2020**, *6*, e19273. doi:10.2196/19273.
7. Qazi, U.; Imran, M.; Ofli, F. GeoCoV19: A Dataset of Hundreds of Millions of Multilingual COVID-19 Tweets with Location Information, 2020, [arXiv:cs.SI/2005.11177].

8. Baran, E.; Dimitrov, D. TweetsCOV19 - A Semantically Annotated Corpus of Tweets About the COVID-19 Pandemic, 2020. doi:10.5281/zenodo.3871753.
9. Lamsal, R. Design and analysis of a large-scale COVID-19 tweets dataset. *Applied Intelligence* **2020**, pp. 1–15.
10. Starbird, K. Disinformation's spread: bots, trolls and all of us. *Nature* **2019**, *571*, 449–450.
11. Loria, S. textblob Documentation. *Release 0.15* **2018**, *2*.
12. Gupta, R.K.; Vishwanath, A.; Yang, Y. Covid-19 twitter dataset with latent topics, sentiments and emotions attributes. *arXiv preprint arXiv:2007.06954* **2020**.
13. Banda, J.M.; Tekumalla, R.; Wang, G.; Yu, J.; Liu, T.; Ding, Y.; Chowell, G. A large-scale COVID-19 Twitter chatter dataset for open scientific research–an international collaboration. *arXiv preprint arXiv:2004.03688* **2020**.
14. Alqurashi, S.; Alhindi, A.; Alanazi, E. Large arabic twitter dataset on covid-19. *arXiv preprint arXiv:2004.04315* **2020**.
15. Feng, Y.; Zhou, W. Is working from home the new norm? an observational study based on a large geo-tagged covid-19 twitter dataset. *arXiv preprint arXiv:2006.08581* **2020**.

Author Contributions

Conceptualization, S.M. and M.K.; methodology, S.M.; software, S.M.; validation, S.M.; formal analysis, S.M.; investigation, S.M. and M.K.; resources, S.M.; data curation, S.M.; writing—original draft preparation, S.M.; writing—review and editing, M.K.; visualization, S.M.; supervision, M.K.; project administration, M.K.; funding acquisition, M.K. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no external funding.

Data Availability Statement

The data presented in this study are openly available at https://zenodo.org/record/4434972.YKA7bJNKhBw, accessed on 15 May 2021.

Conflicts of Interest

The authors declare no conflict of interest.