Applied Intelligence (2021) 51:2790–2804
https://doi.org/10.1007/s10489-020-02029-z

Design and analysis of a large-scale COVID-19 tweets dataset

Rabindra Lamsal[1]

Accepted: 16 October 2020 / Published online: 6 November 2020

Abstract

As of July 17, 2020, more than thirteen million people have been diagnosed with the Novel Coronavirus (COVID-19), and half a million people have already lost their lives due to this infectious disease. The World Health Organization declared the COVID-19 outbreak as a pandemic on March 11, 2020. Since then, social media platforms have experienced an exponential rise in the content related to the pandemic. In the past, Twitter data have been observed to be indispensable in the extraction of situational awareness information relating to any crisis. This paper presents `COV19Tweets Dataset` (Lamsal 2020a), a large-scale Twitter dataset with more than 310 million COVID-19 specific English language tweets and their sentiment scores. The dataset's geo version, the `GeoCOV19Tweets Dataset` (Lamsal 2020b), is also presented. The paper discusses the datasets' design in detail, and the tweets in both the datasets are analyzed. The datasets are released publicly, anticipating that they would contribute to a better understanding of spatial and temporal dimensions of the public discourse related to the ongoing pandemic. As per the stats, the datasets (Lamsal 2020a, 2020b) have been accessed over 74.5k times, collectively.

Keywords Social computing · Crisis computing · Sentiment analysis · Network analysis · Twitter data

1 Introduction

1.1 Social media and crisis events

During a crisis, whether natural or man-made, people tend to spend relatively more time on social media than the normal. As crisis unfolds, social media platforms such as Facebook and Twitter become an active source of information [20] because these platforms break the news faster than official news channels and emergency response agencies [23]. During such events, people usually make informal conversations by sharing their safety status, querying about their loved ones' safety status, and reporting ground level scenarios of the event [11, 20]. This process of continuous creation of conversations on such public platforms leads to accumulating a large amount of socially generated data. The amount of data can range from hundreds of thousands to millions [25]. With proper planning and implementation, social media data can be analyzed and processed to extract situational information that can be further used to derive actionable intelligence for an effective response to the crisis. The situational information can be extremely beneficial for the first responders and decision-makers to develop strategies that would provide a more efficient response to the crisis.

In recent times, the most used social media platforms for informal communications have been Facebook, Twitter, Reddit, etc. Amongst these, Twitter, the microblogging platform, has a well-documented Application Programming Interface (API) for accessing the data (tweets) available on its platform. Therefore, it has become a primary source of information for researchers working on the Social Computing domain. Earlier works [10, 12, 14, 16, 21, 32, 43, 51, 53, 54] have shown that the tweets related to a specific crisis can provide better insights about the event. In the past, millions of tweets specific to crisis events such as the Nepal Earthquake, India Floods, Pakistan Floods, Palestine Conflict, Flight MH370, etc., have been collected and made available [22]. Such Twitter data have been used in designing machine learning models [21, 31, 35] for classifying unseen tweets to various categories such as community needs, volunteering efforts, loss of lives, and infrastructure damages. The classified tweets corpora can

This article belongs to the Topical Collection: *Artificial Intelligence Applications for COVID-19, Detection, Control, Prediction, and Diagnosis*

✉ Rabindra Lamsal
rabindralamsal@outlook.com

[1] School of Computer and Systems Sciences, Jawaharlal Nehru University, New Delhi 110067, India

be (i) trimmed [38] or summarized [36, 40, 41, 50] and sent to the relevant department for further analysis, (ii) used for sketching alert-level heat maps based on the location information contained within the tweet metadata or the tweet body.

Similarly, Twitter data can also be used for identifying the flow of fake news [7, 8, 24, 49]. If miss-information and unverified rumors are identified before they spread out on everyone's news feed, they can be flagged as spam or taken down. Further, in-depth textual analyses of Twitter data can help (i) discover how positively or negatively a geographical region is being textually-verbal towards a crisis, (ii) understand the dissemination processes of information throughout a crisis.

1.2 Novel Coronavirus (COVID-19)

As of July 17, 2020, the number of Novel coronavirus (COVID-19) cases across the world had reached more than thirteen million, and the death toll had crossed half a million [52]. States and countries worldwide are trying their best to contain the spread of the virus by initiating lockdown and even curfew in some regions. As people are bound to work from home, social distancing has become a new normal. With the increase in the number of cases, the pandemic's seriousness has made people more active in social media expression. Multiple terms specific to the pandemic have been trending on social media for months now. Therefore, Twitter data can prove to be a valuable resource for researchers working in the thematic areas of Social Computing, including but not limited to sentiment analysis, topic modeling, behavioral analysis, fact-checking and analytical visualization.

Large-scale datasets are required to train machine learning models or perform any kind of analysis. The knowledge extracted from small datasets and region-specific datasets cannot be generalized because of limitations in the number of tweets and geographical coverage. Therefore, this paper introduces a large-scale COVID-19 specific English language tweets dataset, hereinafter, termed as the `COV19Tweets Dataset`. As of July 17, 2020, the dataset has more than 310 million tweets and is available at IEEE DataPort [30]. The dataset gets a new release every day. The dataset's geo version, the `GeoCOV19Tweets Dataset`, is also made available [29]. As per the stats reported by the IEEE platform, the datasets [29, 30] have been accessed over 74.5k times, collectively, worldwide.

1.3 Organization of the paper

The paper is organized as follows: Section 2 reviews related research works. Section 3 discusses the design methodology of the `COV19Tweets Dataset` and its geo version. Section 4 focuses on the hydration of tweets ID for obtaining full tweet objects. Section 5 presents the analysis and discussions, and Section 6 concludes the paper.

2 Related work

2.1 COVID-19 tweets dataset

Multiple other studies have also been collecting and sharing large-scale datasets to enable research in understanding the public discourse regarding COVID-19. Some of those publicly available datasets are multi-lingual [4, 13, 26, 39], and some are language-specific [3, 18]. Among those datasets, [4, 13, 39] have significantly large numbers of tweets in their collection. [39] provides more than 524 million multi-lingual tweets and also an English version as a secondary dataset. However, with the last update released on May 01, 2020, the dataset [39] does not seem to be getting frequent releases. [4] shares around 490 million multi-lingual tweets alongside the most frequently used terms. [13] provides 302 million multi-lingual tweets, with around 200 million tweets in the English language. However, neither of them [4, 13] have English version releases.

2.1.1 Issues with multi-lingual datasets

First, the volume of English tweets in multi-lingual datasets can become an issue. Twitter sets limits on the number of requests that can be made to its API. Its filtered stream endpoint has a rate limit of 450 requests/15-minutes per app., which is why the maximum number of tweets that can be fetched in 24 hours is just above 4 million. The language breakdown of multi-lingual datasets shows a higher prevalence of English, Spanish, Portuguese, French, and Indonesian languages [4, 13]. Therefore, multi-lingual datasets contain relatively fewer English tweets, unless multiple language-dedicated collections are run and merged later. Second, the size and multi-lingual nature of large-scale datasets can become a concern for researchers who need only the English tweets. For that purpose, the entire dataset must be hydrated and then filtered, which can take multiple weeks.

2.2 Sentiment analysis

Recent studies have done sentiment analysis on different samples of COVID-19 specific Twitter data. A study [1] analyzed 2.8 million COVID-19 specific tweets collected between February 2, 2020, and March 15, 2020, using frequencies of unigrams and bigrams, and performed sentiment analysis and topic modeling to identify Twitter users' interaction rate per topic. Another study [34] examined

tweets collected between January 28, 202, and April 9, 2020, to understand the worldwide trends of emotions (fear, anger, sadness, and joy) and the narratives underlying those emotions during the pandemic. A regional study [33] in Spain performed sentiment analysis on 106,261 conversations collected from various digital platforms, including Twitter and Instagram, during March and April 2020, to examine the impact of risk communications on emotions in Spanish society during the pandemic. In a similar regional study [42] concerning China and Italy, the effect of COVID-19 lockdown on individuals' psychological states was studied using the conversations available on Weibo (for China) and Twitter (for Italy) by analyzing the posts published two weeks before and after the lockdown.

2.3 Network analysis

Multiple studies have performed social network analysis on Twitter data related to the COVID-19 pandemic. A case study [17] examined the propagation of the #FilmYourHospital hashtag using social network analysis techniques to understand whether the hashtag virality was aided by bots or coordination among Twitter users. Another study [2] collected tweets containing the #5GCoronavirus hashtag between March 27, 2020, and April 4, 2020, and performed network analysis to understand the drivers of the 5G COVID-19 conspiracy theory and strategies to deal with such misinformation. A regional study [37] concerning South Korea used network analysis to investigate the information transmission networks and news-sharing behaviors regarding COVID-19 on Twitter. A similar study [27] investigated the relationship between social network size and incivility using the tweets originating from South Korea between February 10, 2020, and February 14, 2020, when the Korean government planned to bring its citizens back from Wuhan.

3 Dataset design

3.1 Data collection

Twitter provides two API types: search API [47] and streaming API [45]. The Standard version of search API can be used to search against the sample of tweets created in the last seven days, while the Premium and Enterprise versions allow developers to access tweets posted in the previous 30 days (30-day endpoint) or from as early as 2006 (Full-archive endpoint) [47]. The streaming API is used for accessing tweets from the real-time Twitter feed [45]. For this study, the streaming API is being used since March 20, 2020.

The original collection of tweets was started on January 27, 2020. The study commenced as an optimization design project to investigate how much social media data volume can be analyzed using minimal computing resources. Twitter's content redistribution policy restricts researchers from sharing tweets data other than tweet IDs, Direct Message IDs and/or User IDs. The original collection did not have tweet IDs. Therefore, tweets collected between January 27, 2020, and March 20, 2020, could not be released to the public. Hence, a fresh collection was started on March 20, 2020.

Figure 1 shows the daily distribution of the tweets in the `COV19Tweets Dataset`. Between March 20, 2020, and April 17, 2020, four keywords, "corona," "#corona," "coronavirus," and "#coronavirus," were used for filtering the Twitter stream. Therefore, the number of tweets captured in that period per day, on average, is around 893k. However, a dedicated collection was started on a Linux-based high-performance CPU-Optimized virtual machine (VM), with additional filtering keywords, after April 18, 2020.

3.1.1 Keywords

As of July 17, 2020, 46 keywords are being tracked for streaming the tweets. The number of keywords has been evolving continuously since the inception of this study. Table 1 gives an overview of the filtering keywords currently in use. As the pandemic grew, a lot of new keywords emerged. In this study, n-grams are analyzed every 2 hours using the recent most 0.5 million tweets to keep track of emerging keywords. Twitter's "worldwide trends" section is also monitored for the same purpose. On May 13, 2020, Twitter also published a list of 564 multi-lingual filtering keywords used in its COVID-19 stream endpoint [44].

The streaming API allows developers to use up to 400 keywords, 5,000 user IDs, and 25 location boxes for filtering the Twitter stream. The keywords are matched against the tokenized text of the body of the tweet. 46 keywords have been identified as filtering rules for extracting COVID-19 specific tweets. User ID filtering was not used. Also, the location box filtering was avoided as the intention was to create a global dataset. Twitter adds a BCP 47[1] language identifier based on the machine-detected language of the tweet body. Since the aim was to pull only the English tweets, the "en" condition was assigned to the language request parameter.

3.2 Infrastructure

The collection of tweets is a small portion of the dataset design. The other tasks include filtration of geo-tagged tweets and computation of sentiment score for each captured tweet, all that in real-time. A dashboard is also required

[1] https://tools.ietf.org/html/bcp47

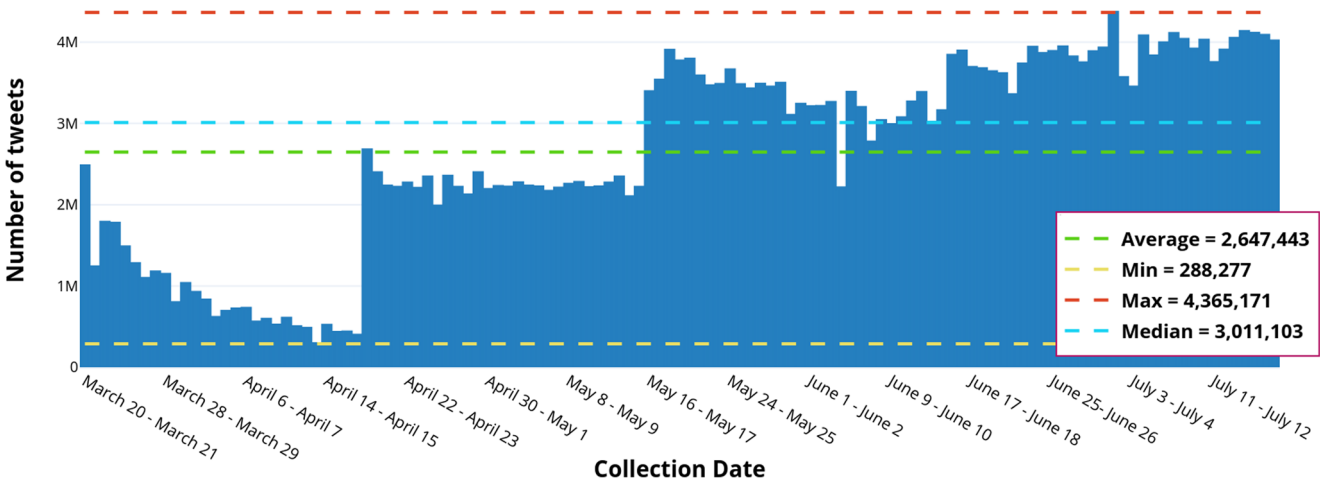


Fig. 1 Daily distribution of tweets in the `COV19Tweets Dataset`

to visualize the information extracted from the collected tweets. A stable internet connection is needed to download the continuously incoming JSON. The computation of sentiment score for each captured tweet requires the VM to constitute powerful enough CPUs to avoid a bottleneck scenario. Every information gathered to this point needs to be stored on a database, which necessitates a disk with excellent performance. Summing up, a cloud-based VM is required to automate all these tasks.

In this study, the VM has to process thousands of tweets every minute. Also, the information extracted from the captured data is to be visualized on an active front-end server that requires plotting of hundreds of thousands of data points. Therefore, a Linux-based compute-optimized hyper-threading VM is used for this study. Table 2 gives an overview of the VM considered in the dataset design. Figure 2a-e shows the resource utilization graphs for various performance parameters of the VM.

A new collection starts between 1000-1100hrs GMT+5:45, every day. Therefore, the CPU usage and average load increase gradually as more and more tweets get captured. The CPU usage graph, in Fig. 2a, shows that the highest percentage of CPU usage at any given time does not exceed 35%. Few Python scripts and libraries, and a web server is actively running in the back-end. The majority of the tasks are CPU intensive; therefore, memory usage does not seem to exceed 35%, as shown in Fig. 2b. Past data show that memory usage exceeds 35% only when the web traffic on the visualization dashboard increases; otherwise, it is usually constant.

The Load average graph, in Fig. 2c, shows that the processors do not operate overcapacity. The three colored lines, magenta, green and purple, represent 1-minute, 5-minute, and 15-minute load average. The Disk I/O graph, in Fig. 2d, interprets the read and write activity of the VM. Saving thousands of tweets information every minute triggers continuous writing activity on the disk. The Disk I/O graph shows that the write speed is around 3.5 MB/s, and the read speed is insignificant. The Bandwidth usage graph, in Fig. 2e, reveals the public bandwidth usage pattern. On average, the VM is receiving a continuous data stream at 3 Mb/s. The VM connects with the backup server's database to download the recent half a million tweets for extracting a list of unigrams and bigrams. A new list is created every 2 hours; therefore, 12 peaks in the Bandwidth usage graph.

Table 1 Overview of the filtering keywords as of July 17, 2020

In use since	Keywords[a]
March 20, 2020	corona, #corona, coronavirus, #coronavirus
April 18, 2020	covid, #covid, covid19, #covid19, covid-19, #covid-19, sarscov2, #sarscov2, sars cov2, sars cov 2, covid_19, #covid_19, #ncov, ncov, #ncov2019, ncov2019, 2019-ncov, #2019-ncov, #2019ncov, 2019ncov
May 16, 2020	pandemic, #pandemic, quarantine, #quarantine, flatten the curve, flattening the curve, #flatteningthecurve, #flattenthecurve, hand sanitizer, #handsanitizer, #lockdown, lockdown, social distancing, #socialdistancing, work from home, #workfromhome, working from home, #workingfromhome, ppe, n95, #ppe, #n95

[a] keyword preceded by a hash sign (#) is a hashtag

Table 2 Overview of the VM

Resource	Description
CPU	Intel(R) Xeon(R) Gold 6140 CPU @ 2.30GHz, Width: 64 bits, 2 vCPUs
Memory	Size: 4GiB
Disk type	Solid State Drive
Bandwidth (based on Speedtest CLI)	Download avg.: 2658.182 Mb/s Upload avg.: 2149.294 Mb/s

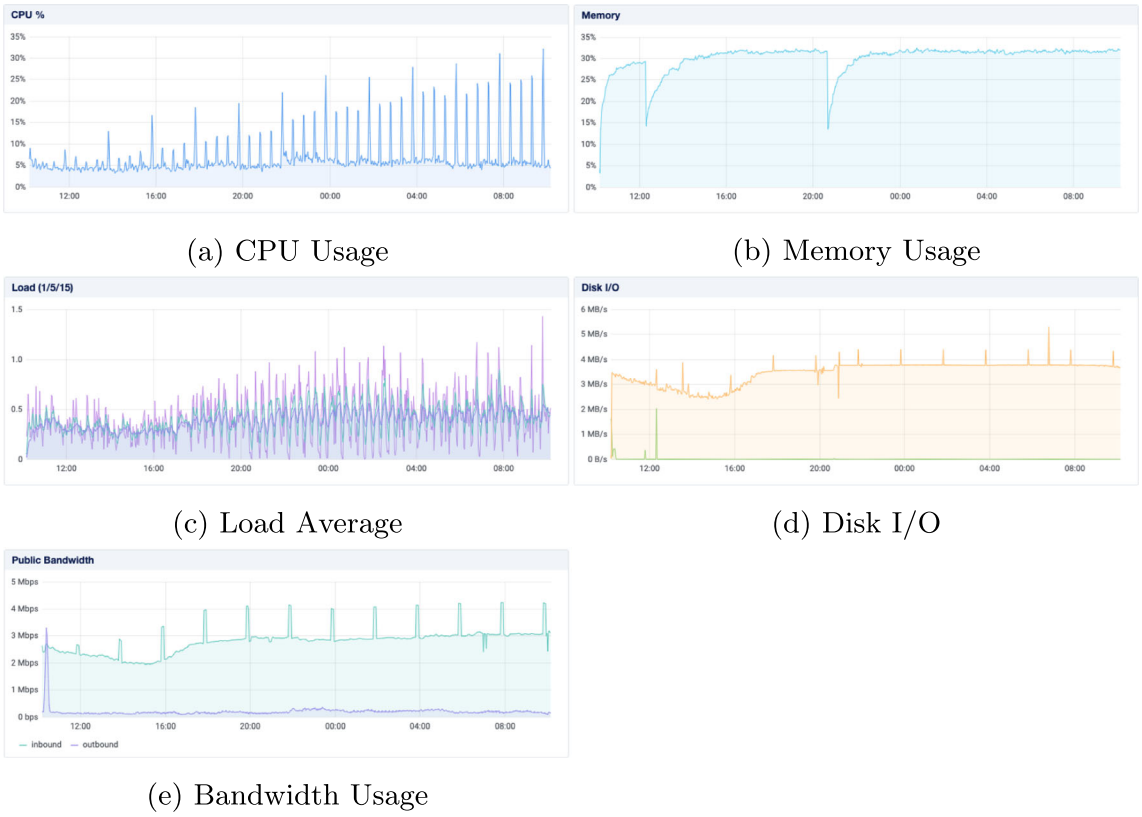
3.3 The sentiment scores

The dataset has two columns: Tweet ID and Sentiment score. During the project's inception, a Long Short-Term Memory (LSTM) deep network was used for computing the sentiment scores. When a new collection was started on March 20, 2020, the LSTM model, which was still in its pre-alpha stage, was replaced by the TextBlob's Sentiment Analysis module. TextBlob is considered among similar libraries since its sentiment analysis model computes the sentiment polarity as a continuous value rather than a category. The sentiment scores are defined in the range [-1,+1]. If a score falls between (0,+1], the tweet is considered to have a Positive sentiment. Similarly, a score in the range [-1, 0) represents a Negative sentiment. And the score "0" denotes a Neutral sentiment. Scores in the extremes of the range [-1,+1] represent strongly Negative sentiment and strongly Positive sentiment, respectively.

Tweets are preprocessed before computing sentiment scores. Hash symbol (#), mention symbol (@), URLs, extra spaces, and paragraph breaks are cleaned. Punctuations, emojis, and numbers are included. Advance-level preprocessing, such as (i) correction of incorrectly spelt words, (ii) conversion of abbreviations to their original forms, are bypassed to avoid analysis bottleneck.

3.4 Filtering geo-tagged tweets

Geotagging is the process of placing location information in a tweet. When a user permits Twitter to access his/her location via an embedded Global Positioning System (GPS), the geo-coordinates data is added to the tweet location



(a) CPU Usage (b) Memory Usage

(c) Load Average (d) Disk I/O

(e) Bandwidth Usage

Fig. 2 Resource utilization graphs for the VM (24 hours)

metadata. This metadata gives access to various Geo Objects [46] such as `"place_type": "city", "name": "Manhattan", "full_name": "Manhattan, NY", "country_code": "US", "country": "United States"` and the bounding box (polygon) of coordinates that encloses the place.

Previous studies have shown that significantly less number of tweets are geo-tagged. A study [5], conducted between 2016-17 in Southampton city, used local and spatial data to show that around 36k tweets out of 5 million had "point" geolocation data. Similarly, in another work [9] done in online health information, it was evident that only 2.02% of tweets were geo-tagged. Further, a multilingual COVID-19 global tweets dataset from CrisisNLP [39] reported having around 0.072% geo-tagged tweets. In this study, the tweets received from the Twitter stream are filtered by applying a condition on the `["coordinates"]` Twitter Object to design the `GeoCOV19Tweets Dataset`.

Algorithm 1 shows the pseudo-code for filtering the geo-tagged tweets. Figure 3 shows the daily distribution of tweets present in the `GeoCOV19Tweets Dataset`. Out of 310 million tweets, 141k tweets (0.045%) were found to be geo-tagged. If the collection after April 18, 2020, is considered, 118k (0.043%) tweets are geo-tagged.

Algorithm 1 Pseudo-code for filtering geo-tagged tweets.

```
while JSON data do
    Load JSON to tweet;
    if tweet["coordinates"]: then
        longitude, latitude =
        data["coordinates"]["coordinates"];
        // Twitter uses
            longitude-latitude order
    end
end
```

3.5 Dataset releases

Twitter's content redistribution policy restricts the sharing of tweet information other than tweet IDs, Direct Message IDs and/or User IDs. Twitter wants researchers to pull fresh data from its platform. It is because users might delete their tweets or make their profile private. Therefore, complying with Twitter's content redistribution policy, only the tweet IDs are released. The dataset is updated every day with the addition of newly collected tweet IDs.

3.5.1 Dataset limitations

First, Twitter allows developers to stream around 1% of all the new public tweets as they happen, via its Streaming API. Therefore, the dataset is a sample of the comprehensive COVID-19 tweets collection Twitter has on its servers. Second, there is a known gap in the dataset. Due to some technical reasons, the tweets collected between March 29, 2020, 1605hrs GMT+5:45, and March 30, 2020, 1400hrs GMT+5:45 could not be retrieved. Third, tweets analysis in a single language increases the risks of missing essential information available in tweets created in other languages [15]. Therefore, the dataset is primarily applicable for understanding the COVID-19 public discourse originating from native English-speaking nations.

4 Using the `COV19Tweets dataset`

Twitter does not allow JSON of the tweets to be shared with third parties; the tweet IDs provided in the `COV19Tweets Dataset` must be hydrated to get the original JSON. This process of extracting the original JSON from the tweet IDs is known as the `hydration of tweets IDs`. There are multiple libraries/applications such as `twarc` (Python

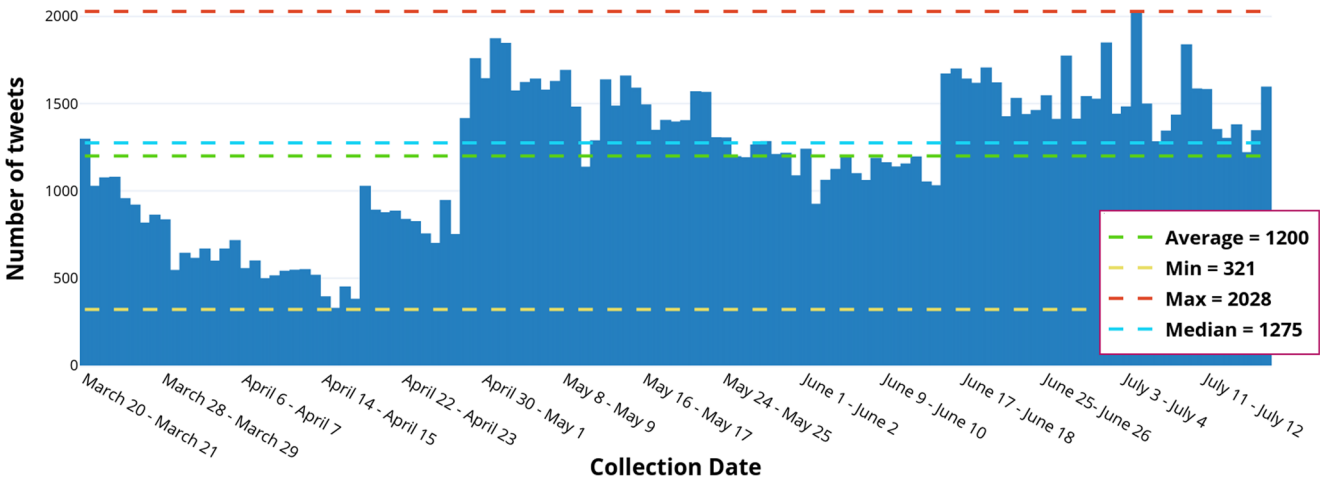


Fig. 3 Daily distribution of tweets in the `GeoCOV19Tweets Dataset`

library) and Hydrator (Desktop application) developed for this purpose. Using the Hydrator application is relatively straightforward; however, working with the twarc library requires basic programming knowledge. Algorithm 2 is the pseudo-code for using twarc to hydrate a list of tweet IDs.

Algorithm 2 Pseudo-code for hydrating a list of tweets IDs.

```
Initialize APIkeys;
t = Twarc(consumer_key, consumer_secret,
access_token, access_token_secret);
for tweet in t.hydrate(open('list_of_ids.csv')) do
    // variable tweet is of type dict
    tweet_id = tweet["id"];
    tweet_text = tweet["full_text"];
    username = tweet["user"]["name"];
    // other tweet information such as
       coordinates, retweet_count,
       follower_count, friends_count, etc,
       can be extracted in similar
       manner
end
```

The tweet data dictionary provides access to a long list of root-level attributes. The root-level attributes, such as user, coordinates, place, entities, etc., further provide multiple child-level attributes. When hydrated, the tweet IDs produce JSON that contains all the root-level and child-level attributes with their values. Twitter's documentation [48] can be referred for more information on the tweet data dictionary.

4.0.2 Filtering tweets originating from a region

The COV19Tweets Dataset has global coverage, and it can also be used to extract tweets originating from a particular region. An implementable solution for this will be to check if a tweet is geo-tagged or has place boundary defined in its data dictionary. If none of these fields are available, the address given on the user's profile can be used. However, Twitter does not validate the profile address field for authentic geo-information. Even addresses such as "Milky Way Galaxy," "Earth," "Land," "My Dream," etc. are accepted entries. A user can also create a tweet from a particular place while having an address of a different one. Therefore, considering user's profile address might not be an effective solution while dealing with location information. Algorithm 3 is the pseudo-code for extracting tweets originating from a region of interest.

Algorithm 3 Pseudo-code for identifying tweets originating from a region of interest.

```
Initialize APIkeys;
t = Twarc(consumer_key, consumer_secret,
access_token, access_token_secret);
for tweet in t.hydrate(open('list_of_ids.csv')) do
    if tweet["coordinates"] then
        if tweet["place"] is not None then
            // lookup the value in
               tweet["place"]["country"] or
               tweet["place"]["country_code"] or
               tweet["place"]["full_name"] to
               check if the tweet is from
               the region of interest
        else
            // some geo-tagged tweets may
               produce NoneType
               tweet["place"]
            geo_long, geo_lat =
            tweet["coordinates"]["coordinates"];
            // convert the coordinates to
               readable address (reverse
               geo-coding) and check if
               the tweet is from the
               region of interest
            // Google Geocoding API
               usage:
               https://maps.googleapis.com
               /maps/api/geocode/json?
               latlng=geo_lat,geo_long&key=
               YOUR_API_KEY
        end
    else if tweet["place"] then
        // lookup the value in
           tweet["place"]["country"] or
           tweet["place"]["country_code"] or
           tweet["place"]["full_name"] to
           check if the tweet is from
           the region of interest
        // place information extracted
           from this block does not
           necessarily mean the tweet
           originated from that
           location
    else
        loc_profile = tweet["user"]["location"];
        // lookup the value in loc_profile
           if the tweet is from the
           region of interest
    end
end
```

5 Analysis & discussions

Tweets received from the Twitter stream can be analyzed for making multiple inferences regarding an event. The tweets collected between April 24, 2020, and July 17, 2020, were considered to generate an overall COVID-19 sentiment trend graph. The sampling time is 10 minutes, which means a combined sentiment score is computed for tweets captured in every 10 minutes. Figure 4 shows the COVID-19 sentiment trend based on public discourse related to the pandemic.

In Fig. 4, there are multiple drops in the average sentiment over the analysis period. In particular, there are fourteen drops where the scores are negative. Among those fourteen drops, seven of the significant drops were studied. The tweets collected in those dates were analyzed to see what particular terms (unigrams and bigrams) were trending. Table 3 lists the most commonly used terms during those seven drops.

The tweets are pre-processed before extracting the unigrams and bigrams. The pre-processing steps include transforming the texts to their lowercases and removing noisy data such as retweet information, URLs, special characters, and stop words [15]. It should be noted that the removal of stop words from the tweet body results in a different set of bigrams. Therefore, the bigrams listed in Table 3 should not be considered the sole representative of the context in which the terms might have been used.

5.1 Network analysis

Next, the `GeoCOV19Tweets Dataset` was used for performing network analysis to extract the underlying relationship between countries and hashtags. Only the hashtags that appear more than ten times in the entire dataset were considered. The dataset resulted in 303,488 number of [country, hashtag] relations from 190 countries and territories, and 5055 unique hashtags. There were 32,474 unique relations when weighted. Finally, the resulting relations were used for generating a network graph, as shown in Fig. 5. The graph shows interesting facts about dataset. The network has a dense block of nodes forming a sphere and multiple sparsely populated nodes connected to the nodes inside the sphere through some relations.

The nodes that are outside the sphere are country-specific hashtags. For illustration, Fig. 6a-d shows the country-specific hashtags for New Zealand, Qatar, Venezuela, and Argentina. The nodes of these countries are outside the sphere because of outliers in their respective sets of hashtags. However, these countries do have connections with the popular hashtags present inside the sphere. The majority of the hashtags in Fig. 6a-d do not relate directly to the pandemic. Therefore, these hashtags can be considered as outliers while designing a set of hashtags for the pandemic.

5.1.1 Communities

The network graph, shown in Fig. 5, is further expanded by a scale factor, as shown in Fig. 7a and b. The network graphs are colored based on the communities detected by a modularity algorithm [6, 28]. The algorithm detected 12 communities in the `GeoCOV19Tweets Dataset`. The `weight='weight'` and `resolution=1.0` parameters were used for the experimentation.

Table 4 gives an overview of the 12 communities identified in the `GeoCOV19Tweets Dataset`. Country names are represented by their ISO codes. Community 0 constitutes 55.56% of the nodes in the network. The number of members in Community 0 was relatively high; therefore, the ISO column for that community lists only the countries that have associations with at least 25 different hashtags. For the remaining communities, all the members are listed.

Communities are formed based on the usage of similar hashtags. The United States has associations with the highest number of different hashtags, it is therefore justified

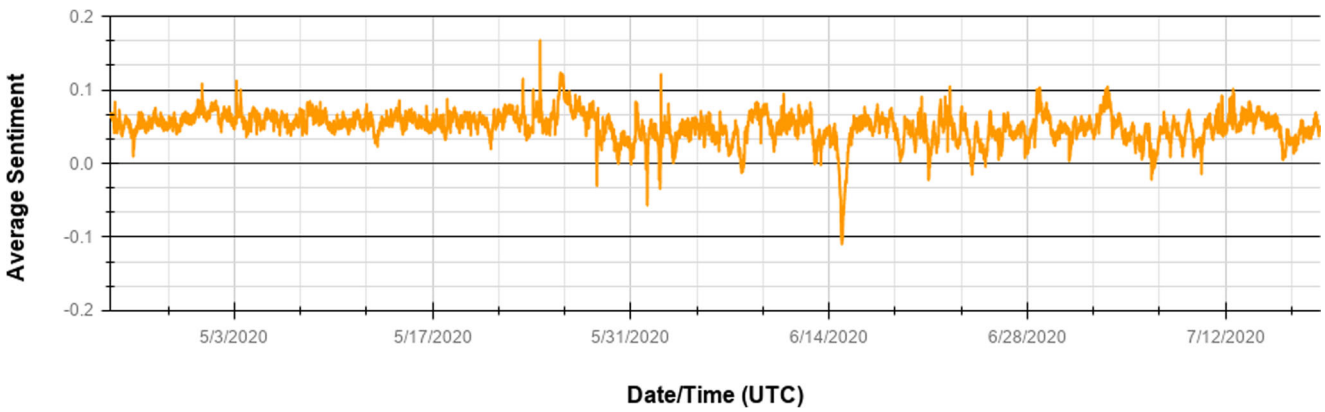


Fig. 4 COVID-19 sentiment trend, since April 24, 2020 to July 17, 2020

Table 3 Trending unigrams and bigrams

Date	score[a]	Unigrams[b]	Bigrams
May 28, 2020	−0.03	deaths, people, trump, pandemic, cases, world, US, virus, health, UK, death, government, china, police	nursing_homes, covid_deaths, bad_gift, tested_positive, gift_china, death_rate, supreme_court, new_york, real_virus, covid_racism
June 01, 2020	−0.05	people, US, health, protests, care, cases, pandemic, home, testing, trump, black, virus, please, masks, curfew, tests	covid_testing, stay_home, testing_centers, impose_curfew, eight_pm, curfue_impose, fighting_covid, peaceful_protests, health_care, enough_masks, masks_ppe
June 14, 2020	−0.11	pandemic, people, children, cases, virus, staff, US, deaths, killed, worst, disease, beat, unbelievable	covid_blacks, latinx_children, unbelievable_asians, systematically_killed, exposed_corona, going_missing, staff_sitting, recovered_covid, worst_disease
June 21, 2020	−0.02	trump, people, pandemic, masks, rally, tulsa, cases, social, distancing, lockdown, died, hospital, mask, call,	wearing_masks, social_distancing, wake_call, mother_died, still_arguing, tested_positive, trump_campaign, tulsa_rally, trump_rally
June 24, 2020	−0.01	pandemic, people, trump, cases, US, testing, lockdown, positive, lindsay, world, social, masks, president	covid_cases, social_distancing, last_year, drunk_driving, lindsay_richardson, tested_positive, wear_mask, america_recovering
July 06, 2020	−0.02	pandemic, people, trump, cases, lockdown, positive, US, virus, wear, social, distancing, mask	social_distancing, got_covid, severe_respiratory, respiratory_cardiovascular, wear_mask, kimberly_guilfoyle, donald_trump
July 10, 2020	−0.01	andemic, coronavirus, people, cases, trump, control, lockdown, US, schools, students, deaths, masks, virus, home, government	control_covid, covid_cases, covid_schools, social_distancing, shake_hands, kneel_bow, hands_hug, vs_right, left_vs

[a] the lowest average sentiment reached on the particular date, [b]excluding the significantly dominating unigrams: COVID, corona, coronavirus and other terms, such as SARS, nCoV, SARS-CoV-2, etc

to find most countries in the same group with the United States. However, other native English-speaking nations such as the United Kingdom and Canada seem to be forming their own communities. This formation of separate communities is because of the differences in their sets of hashtags. For example, the United Kingdom appears to be mostly using "lockdown," "lockdown2020," "isolation," "selfisolation," etc. as hashtags, but the presence of these hashtags in the hashtag set of the United States is limited. The ISO codes for each community in Table 4 are sorted in descending order;

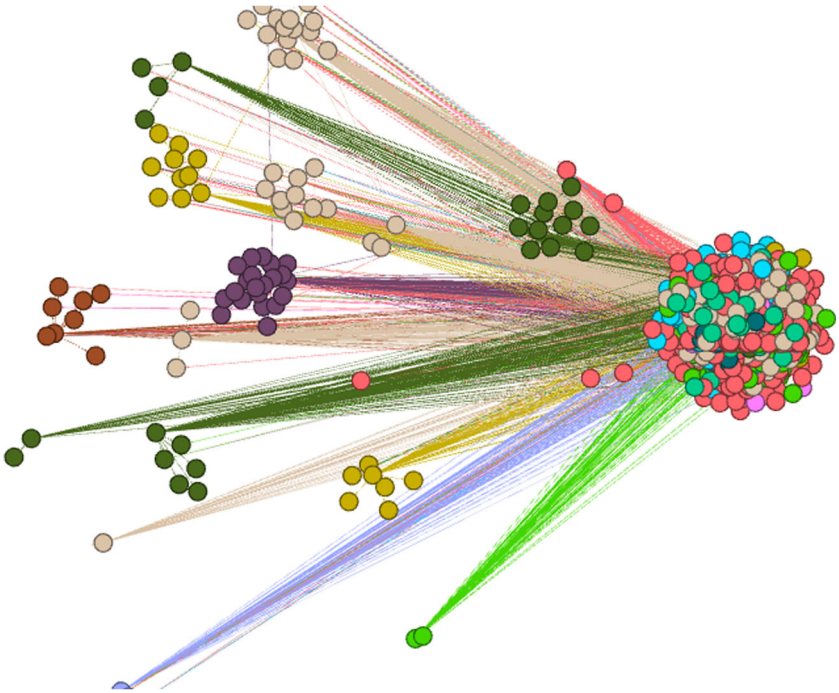
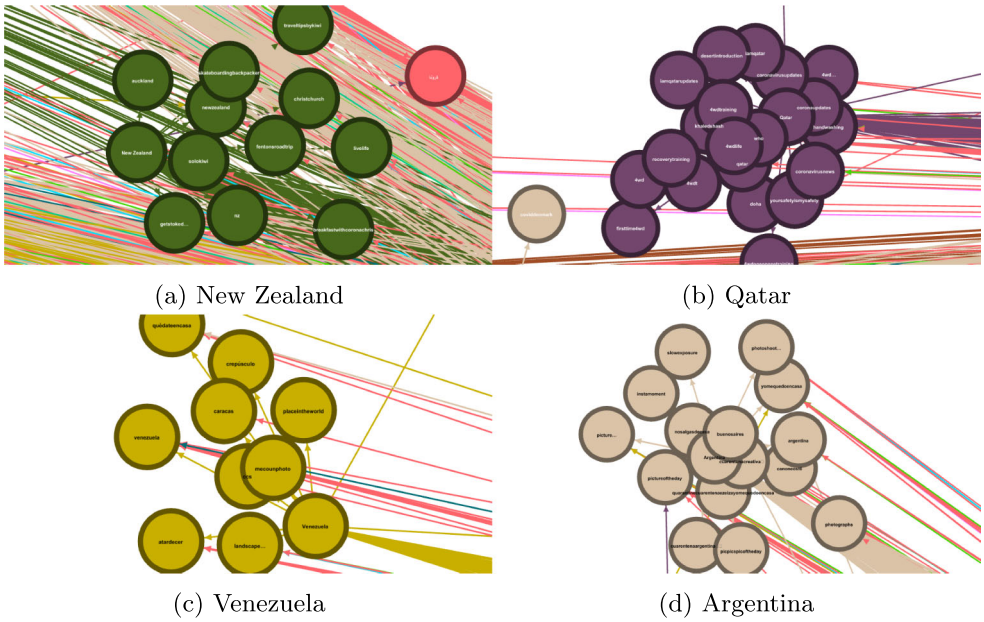


Fig. 5 Network Analysis: Overview of the `GeoCOV19Tweets Dataset`



(a) New Zealand (b) Qatar

(c) Venezuela (d) Argentina

Fig. 6 Country specific outlier hashtags detected using Network Analysis

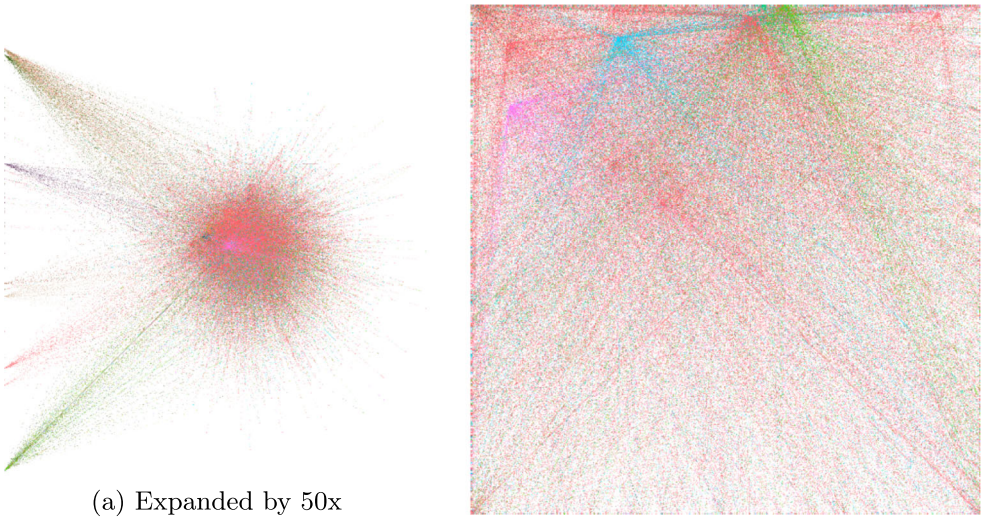
the country associated with the highest number of unique hashtags is mentioned first.

Next, a set of popular hashtags and their communities are identified. Table 5 lists the top 40 commonly used hashtags, their weighted in-degree, and their respective communities. The community for a hashtag in Table 5 means that the hashtag has appeared the most in that particular community.

The [country, hashtag] relations can also be used to trace back a hashtag's usage pattern. The hashtags "flattenthecurve," "itsbetteroutside," "quarantine," "socialdistancing," etc. seem to be first used in the tweets originating from the United States. In the fourth week of March 2020, countries such as the United Kingdom, India, and South Africa experienced their first phase of lockdown. For the same reason, there is an unusual increase in the usage of "lockdown" related hashtags during that period in those countries. It should be noted that a thorough tracing back of hashtag usage would require analysis of tweets collected since December 2019, when the "first case" of COVID-19 was identified [19].

5.2 Sentiment Map

As of July 17, 2020, the number of tweets in the `GeoCOV19Tweets Dataset` is 141,260. The dataset is hydrated to create a country-level distribution of the geotagged tweets, as shown in Table 6. The United States dominates the distribution with the highest number of



(a) Expanded by 50x (b) Expanded by 500x

Fig. 7 Network diagram in Fig. 5 expanded by a scale factor

Table 4 Communities in the `GeoCOV19Tweets dataset`

S No.	C[a]	Color	N[b]	Countries (ISO)[c]
1	0	Medium Red	55.56%	US, AU, NG, ZA, AE, ES, ID, IE, MX, PK, SG, FR, BE, GH, KE, TH, SE, AT, SA, PT, LB, UG, EG, CO, MA, LK, EC, HK, KW, RO, PE, FI, HR, NO, ZW, PA, TZ, VN, BS, PG, HU, BH, CR, BB, OM, SX, RS, TW, BG, DO, ZM, AW, KH, GU, BT, BW, CM, CG, CD, FJ, AQ, SV, AL, ET, JO, UY
2	4	Cyan	17.12%	GB, MV, MK, MU, SK, *SC*, *SY*, *IM*, *CU*, *MO*, *SR*, *GL*, *CK*, *LS*
3	3	Yellow Green	11.55%	IN, TT, *BJ*, *LY*, *TO*
4	2	Blush Pink	4.79%	CA
5	1	Cameo	4.52%	PH, MY, BR, TR, AR, IL, DK, RU, DX, GT, CY, IQ, *AG*, *HN*, *BY*, *TC*, *AI*
6	6	Buddha Gold	2.25%	IT, SI, VE, MC
7	5	Caribbean Green	2.21%	DE, NL, CZ, UA, AO, *GN*
8	9	Pine Green	0.74%	JP, PL
9	8	Fern Frond	0.55%	NZ, NP, MT, IR
10	7	Eggplant	0.42%	QA
11	10	Paarl	0.19%	BM
12	11	Melrose	0.1%	KR

[a] community, [b] percentage of total nodes, [c] italicized ISO codes suggest that those countries have associations with less than 25 different hashtags

geo-tagged tweets followed by the United Kingdom, Canada, and India. During hydration, it was observed that 2.80% of the tweets were either deleted or made private.

The `GeoCOV19Tweets Dataset` has tweets originating from 204 different countries and territories. Around 0.23% of tweets have geo-coordinates information but still produce `NoneType ["place"]` attribute. Such tweets cannot be hydrated to extract place information unless the coordinates are reverse geo-coded. Therefore, the first `if-else` block of Algorithm 3 checks if there is a

Table 5 Top 40 hashtags and their communities

Hashtag	Weighted in-degree	C[a]	Hashtag	Weighted in-degree	C[a]
covid19	31,414	0	isolation	799	4
coronavirus	15,709	0	india	716	3
corona	11,338	0	savetheworld	708	0
lockdown	5,300	4	facemask	704	0
quarantine	5,242	0	workfromhome	655	3
socialdistancing	4,438	0	stayhealthy	634	0
stayhome	4,198	0	savetheworldthanksdoc	617	3
covid	4,074	0	london	568	4
staysafe	3,393	0	health	533	2
pandemic	2,206	0	italy	471	6
billionshieldschallenge	2,129	0	wearamask	459	0
billionshields	1,957	0	fitness	450	4
stayathome	1,675	1	exoworldnow	437	0
faceshield	1,524	0	besafe	435	0
love	1,442	0	newnormal	410	1
quarantinelife	1,323	0	stayhomestaysafe	393	3
mask	1,212	0	selfisolation	391	4
2020	1,192	0	washyourhands	390	0
virus	1,148	0	coronamemes	383	3
lockdown2020	853	4	workingfromhome	364	4

[a] community

Table 6 Distribution of tweets in the `GeoCOV18Tweets Dataset` (top 7)

S No.	Country	# of tweets[a] (n=137,302)
1	United States	60,016 (43.71%)
2	United Kingdom	20,847 (15.18%)
3	Canada	10,688 (7.78%)
4	India	10,082 (7.34%)
5	Nigeria	4,246 (3.09%)
6	Australia	2,893 (2.11%)
7	South Africa	2,824 (2.06%)

[a] as of July 17, 2020, 1010hrs GMT+5:45

requirement for converting geo-coordinates to a human-readable address.

5.2.1 Visualizing the tweets

Next, the geo-tagged tweets were visualized on a map based on their sentiment scores. Figures 8 and 9 are the sentiment maps generated based on the location information extracted from the tweets collected between March 20, 2020, and July 17, 2020. The world view of the COVID-19 sentiment map, in Fig. 8, shows that the majority of the tweets are originating from North America, Europe, and the Indian subcontinent. Interestingly, some tweets are also seen to be originating from countries where the government has banned Twitter. Around 0.26% of the geo-tagged tweets have come from the People's Republic of China, while North Korea does appear on the list, the number is insignificant.

When a region-specific sentiment map, as shown in Fig. 9, is generated, numerous clusters of geo-location points are observed. Such clusters can be a bird's-eye view for the authorities to create first-hand sketches of tentative locations to start for responding to a crisis. For example, the location information extracted from the tweets classified to the "Infrastructure and utilities damage" category can help generate near real-time convex closures of the crisis-hit area. Such convex closures can prove to be beneficial for the first responders (army, police, rescue teams, first-aid volunteers, etc.) to come up with actionable plans. In general, the inferences made from geo-specific data can help (i) understand knowledge gaps, (ii) perform surveillance for prioritizing regions, and (iii) recognize the urgent needs of a population [39].

Understanding the knowledge gaps involves identifying the crisis event-related queries posted by the public on social media. The queries can be anything, a rumor, or even some casual inquiry. Machine learning models can be trained on large-scale tweets corpus for classifying the tweets into multiple informational categories, including a separate class for "queries." Even after the automatic classification, each category still contains hundreds of thousands of tweets conversations, which require further in-depth analysis. Those classified tweets can be summarized to extract concise and important set of conversations. Recent studies have used extractive summarization [41, 50], abstractive summarization [36], and the hybrid approach [40] for summarizing microblogging streams. If the queries are identified and duly answered, the public's tendency to panicking can be settled to some extent.

Further, geo-specific data can assist in surveillance purposes. The social media messages can be monitored actively to identify the messages that report a disease's signs and symptoms. If such messages are detected quite early, an efficient response can be targeted to that particular region. The authorities and decision-makers can come up with effective and actionable plans to minimize possible

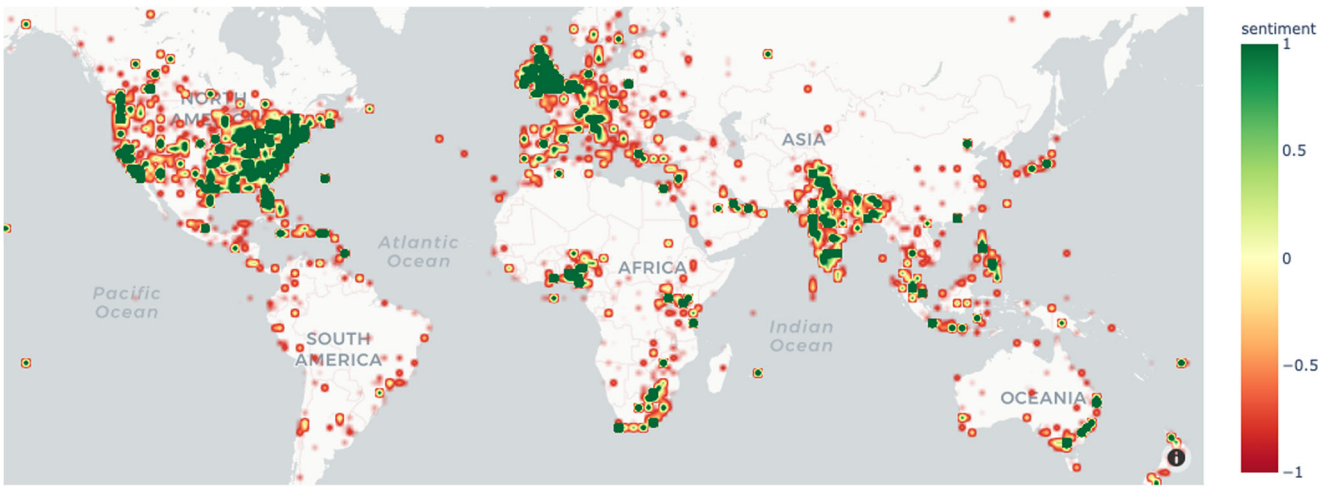
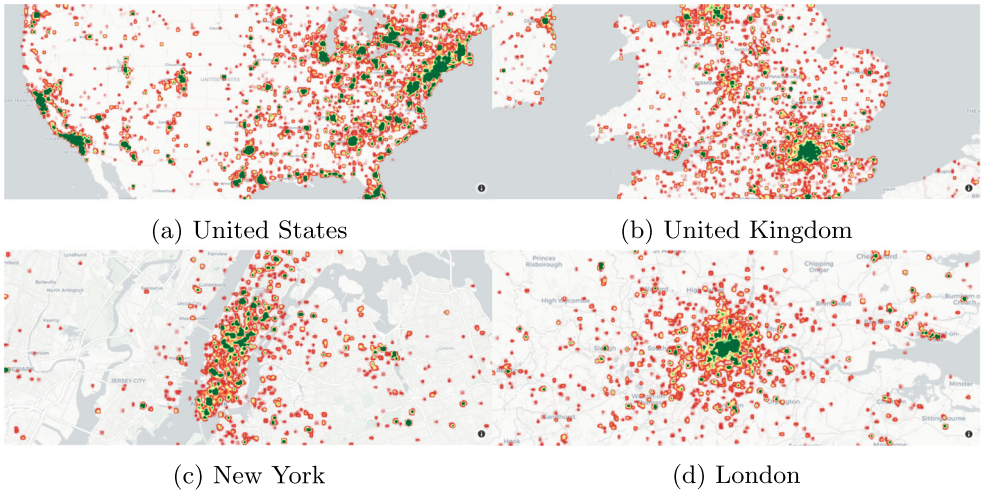


Fig. 8 World view of COVID-19 Sentiment



(a) United States (b) United Kingdom

(c) New York (d) London

Fig. 9 Region-specific view of COVID-19 Sentiment (color scale for this figure is same as of Fig. 8)

future severity. Furthermore, social media messages can also be analyzed to understand the urgent needs of a population. The requirements might include anything related to everyday essentials (shelter, food, water) and health services (medicines, checkups).

The above-discussed research implications fall under the crisis response phase of the disaster management cycle. However, other sub-areas in the Social Computing domain enforce the computational systems to also understand the psychology, and sociology of the affected population/region as part of the crisis recovery phase. The design of such computational systems requires a humongous amount of data for modeling intelligence within them to track the public discourse relating to any event. Therefore, a large-scale Twitter dataset for the COVID-19 pandemic was presented in this paper, hoping that the dataset and its geo version would help researchers working in the Social Computing domain to better understand the COVID-19 discourse.

6 Conclusion

In this paper, a large-scale global Twitter dataset, `COV19Tweets Dataset`, is presented. The dataset contains more than 310 million English language tweets, originating from 204 different countries and territories worldwide, collected over March 20, 2020, and July 17, 2020. Earlier studies have shown that geo-specific social media conversations aid in extracting situational information related to an ongoing crisis event. Therefore, the geo-tagged tweets in the `COV19Tweets Dataset` is filtered to create its geo version, the `GeoCOV19Tweets Dataset`.

Out of 310 million tweets, it was observed that only 141k tweets (0.045%) had "point" location in their metadata. The United States dominates the country-level distribution of the geo-tagged tweets and is followed by the United Kingdom, Canada, and India. Designing a large-scale Twitter dataset requires a reliable VM to fully automate the associated tasks. Five performance metrics (specific to CPU, memory, average load, disk i/o, bandwidth) were analyzed to see how the VM was performing over a period (24 hour). The paper then discussed techniques to hydrate tweet IDs and filter tweets originating from a region of interest.

Next, the `COV19Tweets Dataset` and its geo version were used for sentiment analysis and network analysis. The tweets collected between April 24, 2020, and July 17, 2020, were considered to generate an overall COVID-19 sentiment trend graph. Based on the trend graph, seven significant drops in the average sentiment over the analysis period were studied. Trending unigrams and bigrams on those particular dates were identified. Further, a detailed social network analysis was done on the `GeoCOV19Tweets Dataset` using [country, hashtag] relations. The analysis confirmed the presence of 12 different communities within the dataset. The formation of communities was based on the usage of similar hashtags. Also, a set of popular hashtags and their communities were identified. Furthermore, the `GeoCOV19Tweets Dataset` was used for generating world and region-specific sentiment-based maps, and the research implications of using geo-specific data were briefly outlined.

Acknowledgments The author is grateful to DigitalOcean and Google Cloud for funding the computing resources required for this study.

Compliance with Ethical Standards

Conflict of interests The author declares that there is no conflict of interest.

References

1. Abd-Alrazaq A, Alhuwail D, Househ M, Hamdi M, Shah Z (2020) Top concerns of tweeters during the covid-19 pandemic: infoveillance study. J Med Internet Res 22(4):e19016
2. Ahmed W, Vidal-Alaball J, Downing J, Seguí F. L (2020) Covid-19 and the 5g conspiracy theory: social network analysis of twitter data. J Med Internet Res 22(5):e19458
3. Alqurashi S, Alhindi A, Alanazi E (2020) Large arabic twitter dataset on covid-19. arXiv:2004.04315
4. Banda JM, Tekumalla R, Wang G, Yu J, Liu T, Ding Y, Chowell G (2020) A large-scale covid-19 twitter chatter dataset for open scientific research–an international collaboration. arXiv:2004.03688
5. Bennett NC, Millard DE, Martin D (2018) Assessing twitter geocoding resolution. In: Proceedings of the 10th ACM Conference on Web Science, pp 239–243
6. Blondel VD, Guillaume JL, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. J Stat Mech: Theory Exper 2008(10):P10008
7. Bondielli A, Marcelloni F (2019) A survey on fake news and rumour detection techniques. Inf Sci 497:38–55
8. Bovet A, Makse HA (2019) Influence of fake news in twitter during the 2016 us presidential election. Nat Commun 10(1):1–14
9. Burton SH, Tanner KW, Giraud-Carrier CG, West JH, Barnes MD (2012) "right time, right place" health communication on twitter: value and accuracy of location information. J Med Internet Res 14(6):e156
10. Carley KM, Malik M, Landwehr PM, Pfeffer J, Kowalchuck M (2016) Crowd sourcing disaster management: The complex nature of twitter usage in padang indonesia. Saf Sci 90, 48–61
11. Castillo C (2016) Big crisis data: social media in disasters and time-critical situations. Cambridge University Press
12. Chatfield AT, Scholl HJJ, Brajawidagda U (2013) Tsunami early warnings via twitter in government: Net-savvy citizens' co-production of time-critical public information services. Govern Inf Quart 30(4):377–386
13. 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 Publ Health Surveill 6(2):e19273
14. Cheong M, Lee VC (2011) A microblogging-based approach to terrorism informatics: Exploration and chronicling civilian sentiment and response to terrorism events via twitter. Inf Syst Front 13(1):45–59
15. Dashtipour K, Poria S, Hussain A, Cambria E, Hawalah AY, Gelbukh A, Zhou Q (2016) Multilingual sentiment analysis: state of the art and independent comparison of techniques. Cogn Comput 8(4):757–771
16. Earle P, Guy M, Buckmaster R, Ostrum C, Horvath S, Vaughan A (2010) Omg earthquake! can twitter improve earthquake response? Seismol Res Lett 81(2):246–251
17. Gruzd A, Mai P (2020) Going viral: How a single tweet spawned a covid-19 conspiracy theory on twitter. Big Data Soc 7(2):2053951720938405
18. Haouari F, Hasanain M, Suwaileh R, Elsayed T (2020) Arcov-19: The first arabic covid-19 twitter dataset with propagation networks. arXiv:2004.05861
19. Huang C, Wang Y, Li X, Ren L, Zhao J, Hu Y, Zhang L, Fan G, Xu J, Gu X et al (2020) Clinical features of patients infected with 2019 novel coronavirus in wuhan, china. lancet 395(10223):497–506
20. Imran M, Castillo C, Diaz F, Vieweg S (2015) Processing social media messages in mass emergency: A survey. ACM Comput Surv (CSUR) 47(4):1–38
21. Imran M, Castillo C, Lucas J, Meier P, Vieweg S (2014) Aidr: Artificial intelligence for disaster response. In: Proceedings of the 23rd International Conference on World Wide Web, pp 159–162
22. Imran M, Mitra P, Castillo C (2016) Twitter as a lifeline: Human-annotated twitter corpora for nlp of crisis-related messages. In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). European Language Resources Association (ELRA):Paris, France
23. Imran M, Ofli F, Caragea D, Torralba A (2020) Using ai and social media multimodal content for disaster response and management: Opportunities, challenges, and future directions. Inf Process Manag 57(5):102261. https://doi.org/10.1016/j.ipm.2020.102261
24. Inuwa-Dutse I, Liptrott M, Korkontzelos I (2018) Detection of spam-posting accounts on twitter. Neurocomputing 315:496–511
25. Kalyanam J, Quezada M, Poblete B, Lanckriet G (2016) Prediction and characterization of high-activity events in social media triggered by real-world news. PloS one 11(12):e0166694
26. Kerchner D, Wrubel L Coronavirus tweet ids. harvard dataverse
27. Kim B (2020) Effects of social grooming on incivility in covid-19. Cyberpsychology, Behavior, and Social Networking
28. Lambiotte R, Delvenne JC, Barahona M (2008) Laplacian dynamics and multiscale modular structure in networks. arXiv:0812.1770
29. Lamsal R (2020b) Coronavirus (covid-19) geo-tagged tweets dataset. https://doi.org/10.21227/fpsb-jz61
30. Lamsal R (2020a) Coronavirus (covid-19) tweets dataset. https://doi.org/10.21227/781w-ef42
31. Lamsal R, Kumar TV (2021) Twitter based disaster response using recurrent nets. Int J Sociotechnol Knowl Dev (IJSKD) 14(4)
32. Landwehr PM, Wei W, Kowalchuck M, Carley KM (2016) Using tweets to support disaster planning, warning and response. Saf Sci 90:33–47
33. de Las Heras-Pedrosa C, Sánchez-Núñez P, Peláez J. I (2020) Sentiment analysis and emotion understanding during the covid-19 pandemic in spain and its impact on digital ecosystems. Int J Environ Res Publ Health 17(15):5542
34. Lwin MO, Lu J, Sheldenkar A, Schulz PJ, Shin W, Gupta R, Yang Y (2020) Global sentiments surrounding the covid-19 pandemic on twitter: analysis of twitter trends. JMIR Publ Health Surveill 6(2):e19447
35. Nguyen DT, Al Mannai KA, Joty S, Sajjad H, Imran M, Mitra P (2017) Robust classification of crisis-related data on social networks using convolutional neural networks. In: Eleventh International AAAI Conference on Web and Social Media
36. Olariu A (2014) Efficient online summarization of microblogging streams. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, Volume 2: Short Papers, pp 236–240
37. Park HW, Park S, Chong M (2020) Conversations and medical news frames on twitter: Infodemiological study on covid-19 in south korea. J Med Internet Res 22(5):e18897
38. Purohit H, Hampton A, Shalin VL, Sheth AP, Flach J, Bhatt S (2013) What kind of# conversation is twitter? mining# psycholinguistic cues for emergency coordination. Comput Hum Behav 29(6):2438–2447
39. Qazi U, Imran M, Ofli F (2020) Geocov19: a dataset of hundreds of millions of multilingual covid-19 tweets with location information. SIGSPATIAL Spec 12(1):6–15
40. Rudra K, Goyal P, Ganguly N, Imran M, Mitra P (2019) Summarizing situational tweets in crisis scenarios: An extractive-abstractive approach. IEEE Trans Comput Soc Syst 6(5):981–993
41. Shou L, Wang Z, Chen K, Chen G (2013) Sumblr: continuous summarization of evolving tweet streams. In: Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval, pp 533–542

42. Su Y, Xue J, Liu X, Wu P, Chen J, Chen C, Liu T, Gong W, Zhu T (2020) Examining the impact of covid-19 lockdown in wuhan and lombardy: a psycholinguistic analysis on weibo and twitter. Int J Environ Res Publ Health 17(12):4552
43. Takahashi B, Tandoc Jr EC, Carmichael C (2015) Communicating on twitter during a disaster: An analysis of tweets during typhoon haiyan in the philippines. Comput Hum Behav 50:392–398
44. Twitter: Covid-19 stream (2020). https://developer.twitter.com/en/docs/labs/covid19-stream/filtering-rules
45. Twitter: Filter realtime tweets (2020). https://developer.twitter.com/en/docs/tweets/filter-realtime/overview
46. Twitter: Geo objects (2020). https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/geo-objects
47. Twitter: Standard search api (2020). https://developer.twitter.com/en/docs/tweets/search/overview
48. Twitter: Twitter object (2020). https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/tweet-object
49. Wang B, Zhuang J (2018) Rumor response, debunking response, and decision makings of misinformed twitter users during disasters. Nat Hazards 93(3):1145–1162
50. Wang Z, Shou L, Chen K, Chen G, Mehrotra S (2014) On summarization and timeline generation for evolutionary tweet streams. IEEE Trans Knowl Data Eng 27(5):1301–1315
51. Wang Z, Ye X, Tsou MH (2016) Spatial, temporal, and content analysis of twitter for wildfire hazards. Nat Hazards 83(1):523–540
52. Worldometer: Covid-19 coronavirus pandemic (2020 (accessed July 13, 2020)). https://www.worldometers.info/coronavirus/
53. Zahra K, Imran M, Ostermann FO (2020) Automatic identification of eyewitness messages on twitter during disasters. Inf Process Manag 57(1):102107
54. Zou L, Lam NS, Cai H, Qiang Y (2018) Mining twitter data for improved understanding of disaster resilience. Ann Amer Assoc Geogr 108(5): 1422–1441

Rabindra Lamsal received his BE in Computer Engineering from Kathmandu University and M.Tech in Computer Science and Technology from Jawaharlal Nehru University (JNU). He was also associated with the Special Centre for Disaster Research, JNU, as a Project associate from 2018-19. His areas of research interest are Machine Learning, Natural Language Processing, Social Media Analytics and Social Computing.