Data and Model Biases in Social Media Analyses: A Case Study of COVID-19 Tweets

Yunpeng Zhao, MS[1], Pengfei Yin, MS[1], Yongqiu Li, BS[1], Xing He, MS[1], Jingcheng Du, PhD[2], Cui Tao, PhD[2], Yi Guo, PhD[1], Mattia Prosperi, PhD[1], Pierangelo Veltri, PhD[3], Xi Yang, PhD[1], Yonghui Wu, PhD[1], Jiang Bian, PhD[1a]

[1]University of Florida, Gainesville, Florida, USA; [2]University of Texas Health Science Center at Houston, Houston, Texas, USA; [3]Magna Graecia University of Catanzaro, Catanzaro, Italy

Abstract

During the coronavirus disease pandemic (COVID-19), social media platforms such as Twitter have become a venue for individuals, health professionals, and government agencies to share COVID-19 information. Twitter has been a popular source of data for researchers, especially for public health studies. However, the use of Twitter data for research also has drawbacks and barriers. Biases appear everywhere from data collection methods to modeling approaches, and those biases have not been systematically assessed. In this study, we examined six different data collection methods and three different machine learning (ML) models—commonly used in social media analysis—to assess data collection bias and measure ML models' sensitivity to data collection bias. We showed that (1) publicly available Twitter data collection endpoints with appropriate strategies can collect data that is reasonably representative of the Twitter universe; and (2) careful examinations of ML models' sensitivity to data collection bias are critical.

Introduction

The coronavirus disease (COVID-19) pandemic has put tremendous strain on the society. As of March 9, 2021, more than 29.1 million Americans have been diagnosed with COVID-19 and more than 526,000 have died.[1] Governments worldwide are trying their best to contain the spread of the virus. Preventative measures, such as social distancing, school closures, and work-from-home policies, implemented by national, state, and local governments, have affected the daily routines of billions of people worldwide and forced many activities and social interactions to be moved online.[2,3] Social media platforms are a way for people stay connected during this pandemic. Individuals are increasingly sharing a large amount of personal health information, including their COVID-19-related sentiments and comments. Officials such as health organizations and government agencies have used social media to share COVID-related policies, progress of vaccine development, and Q&A towards COVID-19 issues to help the public stay safe and informed.[4] These social media data provide unique insights into public health events.

Among the popular social media platforms, Twitter, initially a microblogging platform, has well-constructed Application Programming Interfaces (APIs) for obtaining the data that are publicly available. Therefore, it has become a popular source of social media data for researchers. In the short time since the pandemic began, Twitter has been used to study various topics around COVID-19. For example, Kouzy *et al.* (2020) manually identified and quantified the magnitude of misinformation that is being spread on Twitter regarding the COVID-19 pandemic which can served as an early warning for unexpected events.[5] Xue *et al.* (2020) used latent Dirichlet allocation (LDA) to identify popular topics and sentiments from 4 million COVID-19 tweets.[6] Mackey *et al.* (2020) used the Biterm topic modeling to identify individual reports of COVID-19-related symptoms, testing, and recoveries that appeared on Twitter.[7]

Twitter provides new opportunities for health-related research. However, the use of Twitter data for research also has drawbacks and difficulties; the potential for bias appears at every stage of the process, from data collection to modeling. Tools for collecting social media data often result in biased samples. For example, Twitter APIs only return a subset of the tweets from Twitter's data warehouse; but the relative size of a subset as a proportions of the whole is unknown, as are the sampling strategies used by Twitter to produce the subsets.[8] Further, social media data (e.g., tweets) are mostly unstructured free-text data. To study health information on Twitter, researchers often use inference models such as machine learning (ML) and topic modeling methods to process and identify insights from these free-text tweets. However, supervised ML models (e.g., models used to identify genuine laypeople discussions from health-related discussions on Twitter[9–11]) require annotated training samples. If ML models are trained on

[a] Corresponding: Jiang Bian, PhD; bianjiang@ufl.edu

potentially biased subsets of data, it is not clear how well those models will perform when analyzing other samples from the whole datasets. Biases of social media data and analysis methods have yet to be rigorously addressed, particularly in public health surveillance studies such as those for COVID-19. Overlooking these biases in health-related social media studies can lead to wrong or inappropriate results with severe unintended consequences.

Thus, in this study, we used six different data collection methods available to collect COVID-19-related Twitter data and developed three machine learning models. This study had two primary aims, listed below with three corresponding research questions (RQs):

Aim 1: Assess data collection bias.

- **RQ1**: What proportions of each dataset data collection method returns?
- **RQ2**: How representative of the data collected with each of the 5 data collection methods to the gold standard dataset (i.e., from the Twitter full archive)?

Aim 2: Measure ML models' sensitivities to data collection bias.

- **RQ3**: How does data collection bias (i.e., models trained on different subsets) impact models' performance when applied on other subsets?

Methods

Figure 1 shows the overall process of our study, where we (1) collected COVID-19 related tweets using 6 different data collection methods, (2) estimated the data collection biases in two-fold: i) assessing the overlapping portions between each pair of the six datasets, and ii) assessing the data representativeness using rank correlation based on the top keywords comparing each of the other 5 data collection methods against the full archive data, and (3) explored ML models' sensitivity to data bias in terms of model performance by training and testing prediction models on samples selected from different datasets.

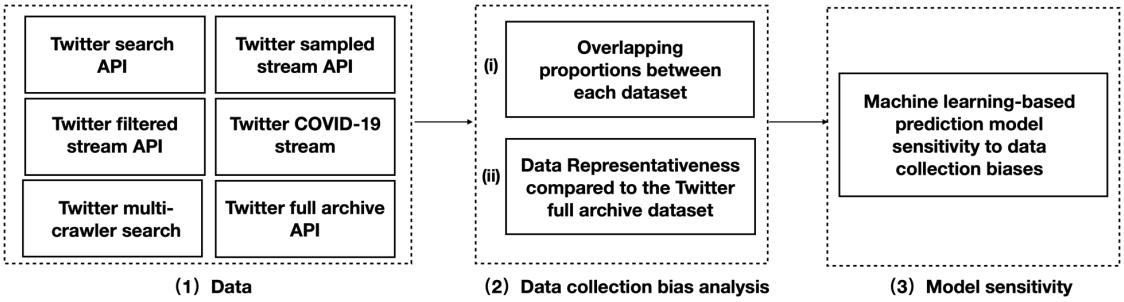

Figure 1. The overall study analysis workflow.

Data collection

From February 21st, 2020 to May 1st, 2020, we collected COVID-19-related tweets using three different Twitter APIs: (1) Twitter search API (i.e. "*GET search/tweets*"),[12] (2) Twitter sampled stream API (i.e. "*GET statuses/sample*"),[13] and (3) Twitter filtered stream API (i.e. "*POST statuses/filter*"[3] using a list of keywords (e.g. *"#coronavirus*" and "*covid*").[14] The list of keywords was developed through a snowball sampling process, where we started with a list of seed keywords collected from online information sources such as news sites and Wikipedia. We then iteratively queried sample tweets from the Twitter website using these keywords and manually reviewed the content of the tweets to discover new COVID-19-related keywords (i.e., words that co-occur with one of the existing keywords but that were not in the existing keyword list) until no new keywords were found. Through this process, we initially found 36 COVID-19-related keywords initially.

As the COVID-19 pandemic progressed, using a similar process, we further extended our 36 keywords to a total of 86 keywords based on a new round of sampling of relevant tweets and Google search results. Since Twitter APIs have rate limits[4] (e.g. for the search AP, only 450 requests are allowed in a 15-minute time window for each crawler and restricted by IP), we split the 86 keywords into 4 groups and used 4 crawlers making 4 sets of Twitter search API requests separately (a.k.a "*Twitter multi-search*" in our experiments) from May 14, 2020 to July 17, 2020.

On April 29, 2020, Twitter released a new streaming interface that is designed for COVID-19 data collection (a.k.a "*Twitter COVID-19 stream*" in our experiment).[15] This COVID-19 endpoints allow approved developers to access COVID-19 related tweets across languages. Nevertheless, the exact mechanism of how Twitter decides that a tweet related to COVID-19 is unclear. Thus, to make this COVID-19 streaming dataset comparable to our other Twitter

data collection methods, we filtered the dataset using the same set of keywords and the same time range as in our multi-crawler data collection method.

In January 2021, Twitter made its full archive data available to academic researchers; however, the full archive APIs are still constrained by the rate limits.[16,17] Ideally, the full archive dataset should contain all tweets of the Twitter universe and thus can be used as the gold-standard dataset for calculating the data coverages of the different data collection methods described above. Because of the rate limits, we collected 200 random samples of a full-minute of tweets from the "*Twitter full archive*" based on the corresponding keywords and the time ranges that we used for the other data collection methods accordingly.

In summary, we used six Twitter data collection methods resulting in six different datasets: (1) "*Twitter search*", (2) "*Twitter filtered stream*", (3) "*Twitter sampled stream"*, (4) "*Twitter COVID-19 stream*", (5) "*Twitter multi-crawler search*", and (6) "*Twitter full archive*". **Figure 2 (A)** shows the time range of each dataset; and as shown in **Figure 2 (B)**, we divided the datasets into two groups based on the data collection time ranges. **Group 1** includes four datasets (i.e. "*Twitter search*", "*Twitter filtered stream*", "*Twitter sampled stream*", and "*Twitter full archive*") collected from February 21, 2020 to May 1, 2020; and Group 2 includes three datasets (i.e. "*Twitter COVID-19 stream*", "*Twitter multi-crawler search*", and "*Twitter full archive*") collected from May 14, 2020 to July 17, 2020. As shown in **Figure 2 (B)**, ideally, each of the other Twitter data collection methods returns a subset of the "*Twitter full archive*" dataset, while overlaps with the other datasets collected from the same time period and with the same list of keywords. However, since we collected the full archive data in January 2021, tweets that have been deleted or from accounts that have been suspended were no longer accessible, resulting in other datasets having tweets not in the full archive. The eight circles in **Figure 2 (B)** represent eight random samples that we used to train our ML models. The results of the training sample annotations and associated ML model performance will be detailed in the result section.

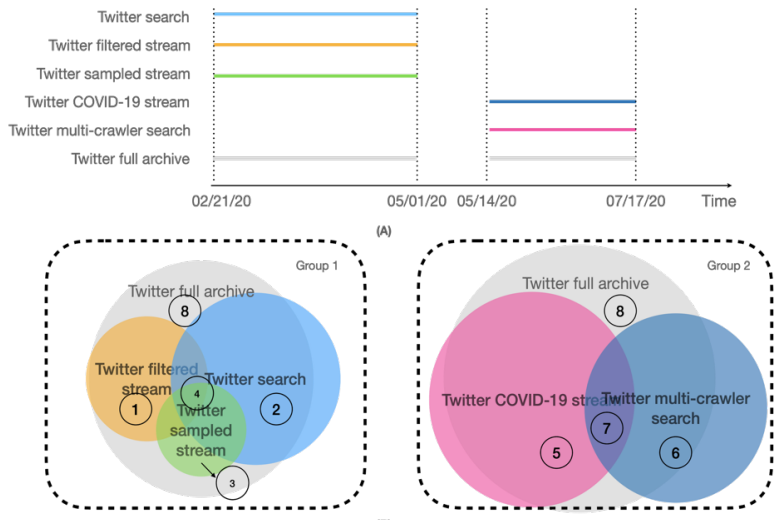

Figure 2. The 6 different Twitter datasets collected by different data collection methods and their corresponding data collection time periods.

Bias in social media data collection methods

The underlying mechanisms of the Twitter APIs' sample selection strategies are unknown. We first assessed the data collection bias in terms of overlaps among the datasets within the two groups (as shown in **Figure 2 (B)**). We then compared the overlapping of tweets by considering the top relevant keywords and measuring the Kendall correlations for each dataset against the full archive dataset.

Originally, we thought that the Twitter full archive API could reliably be used as the gold-standard dataset; however, since we collected the full archive data in January 2021, tweets that had been deleted and tweets from suspended Twitter accounts were no longer available. Thus, when we compared each dataset with the full archive, these tweets had to be removed from our calculations.

In addition to comparing the first five data collection methods with the "*Twitter full archive*" dataset over the span of a minute, we also measured the overlapping among the first five datasets within additional time spans, including a minute, an hour, a day, and a week. To do so, we first randomly selected 100 random time periods at the lengths of a minute, hour, day, and week. We then calculated the overlapping proportions of the datasets generated from each data collection method and reported the mean overlapping proportions and associated confidence intervals across 100 random samples. For Group 1, we calculated the overlapping proportions among the datasets using the total of the first three datasets (i.e., "*Twitter search*", "*Twitter filtered stream*", and "*Twitter sampled stream*"); and for Group 2, the denominator consisted of the total of the "*Twitter multi-crawler search*" and "*Twitter COVID-19 stream*" datasets.

In addition to calculating the global overlapping proportions among these five datasets, we also aimed to answer RQ2:

"*how representative of the data collected with each of the 5 data collection methods to the gold standard dataset (i.e. from the Twitter full archive)?*". To make a fair comparison, we first identified the top 10 keywords of each dataset. We then measured the overlapping proportions between each of the first five datasets and the gold-standard dataset, only considering tweets containing the top 10 keywords. To measure the representativeness, we performed Kendall correlations for each comparison.

Machine learning models' sensitivity to data bias

To test machine learning modes' sensitivity to data bias, we considered a prediction task that classifies task, based on our past work,[10] that classifiers each tweet into two categories: promotional information and consumer discussions.

Training sample selection. We randomly selected 600 English tweets from eight different sampling points (4,800 total tweets) as shown in **Figure 2 (B)**. Those 4,800 tweets were manually sorted into the 3 groups (i.e., irrelevant, promotional, and consumer discussions) by three annotators (kappa = 0.73). Even if a tweet contains keywords related to COVID-19, the tweet may not be relevant or meaningful (e.g., *"#SIRCSWORLDNEWS #Coronavirus #CoronavirusOutbreak -- BREAKING - <URL>"*); thus, we categorized those tweets as irrelevant. Within the relevant tweets, we further categorized those tweets into promotional information (e.g., from a health organization "*CDC Denies Delaying Testing of California Coronavirus Patient; Over 100 CA Hospital Workers in Home Quarantine <URL>*") and consumer discussions (e.g., "*My cousin just died of Coronavirus <URL>*").

Tweet text preprocessing. To build the classifiers, we first preprocessed the sampled tweets following the preprocessing steps used by GloVe:[18] (1) removed hyperlinks, (2) removed mentions, (3) replaced hashtags into English words with hashtag sign (e.g., convert "#COVID" to "<hashtag> COVID"), and (4) replaced all emojis, URLs, and mentions (e.g., @username) with signs of "<emojis", "<url>", and "<user>" respectively.

Machine learning model sensitivity to data bias. We explored three commonly used classification algorithms: convolutional neural networks (CNN), random forest (RF), and gradient boosting trees (GBT).[19–21] We implemented the CNN models in Keras on top of the Tensorflow framework. We initialized the embedding layer with the GloVe pre-trained 100-dimension Twitter word embeddings. We implemented the RF and GBT via the scikit-learn library and used the Term Frequency-Inverse Document Frequency (TF-IDF) scheme to convert each tweet into a feature vector.

Regarding our research question about "*model sensitivity to data collection bias*", our hypotheses were: (1) prediction models that were trained on samples selected from one dataset may not achieve consistent performance when they are applied on data from other datasets; and (2) prediction models trained on samples selected from the "*Twitter full archive*" dataset should achieve relatively higher performance and higher consistency compared to the models trained by samples from the other collection methods, since these are theoretically samples of the full archive.

To test our hypotheses, we trained each model 10 times on each of the eight annotated samples and then tested the model's performance on the other samples. The performances were measured in terms of mean F-1 score and the 95% confidence intervals (CI) were reported across 10 runs.

Results

Data collection

In total, we collected more than 750 million tweets using six different Twitter APIs across two different time periods.

Table 1. Tweet counts by data collection method, corresponding time range, and number of keywords used.

Data collection methods	Number of tweets	Time range	Number of keywords
Twitter search API	200,423,651	02/21/20 - 05/01/20	36
Twitter filtered stream API	108,987,452	02/21/20 - 05/01/20	36
Twitter sampled stream API	3,145,428	02/21/20 - 05/01/20	36
Twitter multi-crawler search	253,996,071	05/14/20 - 07/17/20	86
Twitter COVID-19 stream	183,815,527	05/14/20 - 07/17/20	86
Twitter full archive	298,040	02/21/20 - 05/01/20	36
	280,342	05/14/20 - 07/17/20	86

Data collection bias analysis

RQ1: What proportions of each dataset data collection method returns?

We randomly selected 100 random time periods at the minute, hour, day, and week scales within each of the two dataset groups. We first compared each dataset against the gold standard dataset at the minute scale by calculating

the overlapping proportion between each dataset and the "*Twitter full archive*" dataset. The denominator used for this comparison was the total number of tweets in the "*Twitter full archive*" dataset. We then compared the overlapping portions of the datasets within each group, using the combination of all available datasets within each group as the denominator. We also identified the top 10 keywords of each dataset as shown in **Table 2**. Except the "*Twitter full archive*" data in Group 1, the other datasets collected at the same time period have the same top 10 keywords within its group. The "*Twitter full archive*" data in Group 1 has 9 keywords that are the same as the other three datasets within that same group, with the exception of *"#viruscorona*" which only exists in the top 10 list of the other sample datasets.

Table2. Top 10 keywords and corresponding number of tweets by datasets.

Group	Datasets	Top 10 keywords
Group 1 (02/21/20 - 05/01/20)	Twitter search	**coronavirus** (n=72,165,676), **covid-19** (n=24,390,634), **#coronavirus** (n=21,836,606), **novel coronavirus** (n=550,429), **#covid-19** (n=503,500), **ncov** (n=315,694), **#coronaoutbreak** (n=243,728), **#wuhanvirus** (n=220,353), **19-ncov** (n=64,158), **#viruscorona** (n=60,339)
	Twitter filtered stream	**coronavirus** (n=77.382,453), **#coronavirus** (n=22,722,083), **covid-19** (n=21,630,298), **#covid-19** (n=524,053), **novel coronavirus** (n=455,227), **#coronaoutbreak** (n=383,747), **ncov** (n=376,125), **#viruscorona** (n=134,011), **#wuhanvirus** (n=85,157), **19-ncov** (n=49,960)
	Twitter sampled stream	**coronavirus** (n=1,273,468), **covid-19** (n=434,379), **#coronavirus** (n=352,080), **ncov** (n=13,598), **#covid-19** (n=8,477), **#coronaoutbreak** (n=6,133), **novel coronavirus** (n=5,436), **#wuhanvirus** (n=4,322), **#viruscorona** (n=1,257), **19-ncov** (n=430)
	Twitter full archive	**coronavirus** (n=188,875), **covid-19** (n=75,789), **#coronavirus** (n=62,469), **#covid-19** (n=1,919), **novel coronavirus** (n=1,848), **#coronaoutbreak** (n=1,076), **ncov** (n=883), **19-ncov** (n=321), 2019-ncov (n=319), **#wuhanvirus** (n=273)
Group 2 (05/14/20 - 07/17/20)	Twitter COVID-19 stream	**covid** (n=111,744,558), **corona** (n=54,737,140), **covid-19** (n=41,067,636), **coronavirus** (n=36,596,671), **#covid** (n=25,918,779), **pandemic** (n=24,241,556), **#corona** (n=9,041,074), **#coronavirus** (n=7,389,182), **corona virus** (n=1,707,980), **outbreak** (n=1,495,608)
	Twitter multi-crawler search	**covid** (n=139,127,350), **corona** (n=78,295,697), **covid-19** (n=65,088,306), **coronavirus** (n=55,843,235), **pandemic** (n=44,196,324), **#covid** (n=23,252,190), **#corona** (n=15,271,233), **#coronavirus** (n=13,063,092), **outbreak** (n=5,00,215), **corona virus** (n=2,102,519)
	Twitter full archive	**covid** (n=93,294), **covid-19** (n=50,136), **corona** (n=49,598), **coronavirus** (n=34,387), **pandemic** (n=25,372), **#covid** (n=18,967), **#corona** (n=18,503), **#coronavirus** (n=15,496), **outbreak** (n=2,133), **corona virus** (n=1,376)

Table 3 (considering all the tweets in each dataset) and **Table 4** (considering tweets that contain the top 10 keywords) show the mean number of tweets and 95% confidence intervals across the 100 random samples at each time scale. The "*Twitter COVID-19 stream*" and "*Twitter multi-crawler*" search APIs collected higher volumes of tweets compared with "*Twitter search"*, "*Twitter filtered stream"*, and "*Twitter sampled stream*" APIs at each time scale.

Table 3. Tweet counts including all the tweets in each dataset at each time scale.

	Scale	Twitter search	Twitter filtered stream	Twitter sampled stream	Twitter full archive	Twitter COVID-19 stream	Twitter multi-crawler search	Twitter full archive
All the tweets of each dataset		**Group 1 (02/21/20 - 05/01/20)**				**Group 2 (05/14/20 - 07/17/20)**		
	Minute (Overlapping with Archive)	1,325 (1,132, 1,518)	1,232 (1,192, 1,272)	21 (19, 23)	2,980 (2,716, 3,244)	1,790 (1,675, 1,906)	2,141 (2,015, 2,266)	2,831 (2,671, 2,991)
	Minute	2,416 (183, 5,118)	1,854 (1,191, 2,220)	48 (16, 97)	NA	2705 (2534, 2876)	2880 (2709, 3050)	NA
	Hour	142,004 (12,894, 286,153)	82,779 (52,420, 100,777)	2,629 (1,053, 5,159)	NA	171,278 (158,370, 184,187)	187,773 (173,357, 202,188)	NA
	Day	3007246 (38,773, 5,030,460)	1,869,184 (1,272,981, 2,205,746)	50,407 (43,213, 57,601)	NA	3,907,443 (3,719,401, 4,095,485)	4,116,652 (3826724, 4406580)	NA
	Week	18,844,663 (18,138,284, 26,844,663)	15142006 (11503525, 26142006)	400787.94 (161826.306, 560712)	NA	25,342,161 (220,159,22, 28668400)	26,712,453 (25,987,015, 30,231,051)	NA

Table 4. Tweet counts and 95% confidence intervals across the 100 random samples at each time scale.

	Scale	Twitter search	Twitter filtered stream	Twitter sampled stream	Twitter full archive	Twitter COVID-19 stream	Twitter multi-crawler search	Twitter full archive
Tweets that contain the top 10 keywords		**Group 1 (02/21/20 - 05/01/20)**				**Group 2 (05/14/20 - 07/17/20)**		
	Minute (Overlapping with Archive)	884 (741, 1,027)	943 (912, 973)	18 (16, 20)	1,856 (1,718, 1,994)	705 (558, 656)	902 (842, 962)	915 (856, 974)
	Minute	1,753 (145, 4,249)	1,674 (856, 2,112)	29 (7, 53)	NA	2,386 (1,656, 3,116)	2,666 (2,496, 2,835)	NA
	Hour	59,906 (55,148, 64,664)	50,389 (42,257, 58,521)	1120 (980, 1259)	59,906 (55,148, 64,664)	143,564 (140,646, 146,483)	162,299 (159,636, 164,962)	NA
	Day	2,413,103 (29,506, 4,372,090)	1,714,022 (1,173,389, 2,081,265)	50,407 (43,213, 57,601)	NA	3,405,842 (3,156,694, 3,654,991)	3,850,300 (3665628, 4034972)	NA
	Week	14,731,671 (7,281,210, 20,093,600)	11,301,472 (9,345,513, 13,319,242)	267,235 (207,605, 320,056)	NA	22,819,147 (17,544,694, 28,093,600)	24,496,456 (21,222,624, 27,770,288)	NA

Table 5 shows the overlapping proportions of each pair of datasets from Group 1. As shown in **Table 5**, the "*Twitter search*" dataset captured most of the relevant tweets (i.e., a larger portion than the others). The overlapping between the "*Twitter search*" dataset and the "*Twitter filtered stream*" are slightly increased at all time scales, comparing only the tweets with the top 10 keywords.

Table 5. Overlapping data among datasets in Group 1.

Scale	Twitter search	Twitter filtered stream	Twitter sampled stream	Twitter search vs. Twitter filtered stream	Twitter search vs. Twitter sampled stream	Twitter filtered stream vs. Twitter sampled stream	Overlapping across the three datasets
Comparison of all tweets of each dataset							
Minute[a]	52.3% (45.1%, 59.6%)	48.2% (44.7%, 51.6%)	0.8% (0.7%, 0.8%)	28.6% (23.4%, 33.3%)	0.4% (0.3%, 0.5%)	0.6% (0.6%, 0.7%)	0.4% (0.3%,0.4%)
Minute[b]	72.7% (49.7%, 95.7%)	64.5% (31.8%, 97.2%)	1.6% (1.3%, 1.9%)	37.7% (23.1%, 52.4%)	0.6% (0.2%, 1.0%)	0.6% (0.5%, 0.8%)	0.6% (0.1%, 1.0%)
Hour[b]	72.9% (51.2%, 94.6%)	54.7% (17.8, 91.6%)	1.8% (0.9%, 2.7%)	28.4% (9.0%, 47.8%)	0.5% (0.3%, 0.7%)	0.6% (0.4%, 0.7%)	0.4% (0.1%, 0.7%)
Day[b]	70.5% (60.1%, 80.9%)	53.4% (28.2%, 78.6%)	1.9% (1.0%, 2.9%)	25.6% (8.4%, 42.8%)	0.5% (0.3%, 0.7%)	0.7% (0.5%, 0.7%)	0.4% (0.1%, 0.7%)
Week[b]	70.4% (65.6%, 75.2%)	54.0% (37.6, 70.4%)	1.9% (1.3%, 2.5%)	27.5% (16%, 39%)	0.5% (0.2%, 0.7%)	0.6% (0.4%, 0.7%)	0.4% (0.2%, 0.5%)
Comparison of the tweets with the top 10 keywords of each dataset							
Minute[a]	57.1% (53.4%, 60.9%)	52.1% (44.6% 59.5%)	1.0% (0.9%, 1.0%)	31.9% (26.3%, 37.6%)	0.5% (0.4%, 0.6%)	0.9% (0.9%, 1.0%)	0.5% (0.4%, 0.6%)
Minute[b]	74.4% (51.7%, 97.1%)	69.4% (42.7%, 96.1%)	1.2% (0.6%, 2.4%)	44.1% (3.1%, 85.4%)	0.6% (0, 1.0%)	1.1% (0.5%, 1.8%)	0.5% (0, 0.1%)
Hour[b]	77.4% (72.5%, 82.2%)	52.9% (47.9%, 58.0%)	1.3% (1.2%, 1.5%)	30.3% (26.4%, 34.3%)	0.5% (0.4%, 0.5%)	0.8% (0.8%, 0.9%)	0.3% (0.3%, 0.4%)
Day[b]	68.8% (42.5%, 95.1%)	58.4% (49.6%, 67.2%)	1.4% (0.7%, 2.1%)	27.7% (15.3%, 40.1%)	0.5% (0.3%, 0.7%)	0.8% (0.5%, 1.6%)	0.3% (0, 0.5%)
Week[b]	71.8% (57.0%, 86.6%)	57.5% (50.1%, 64.9%)	1.3% (0.7%, 1.9%)	28.6%s (17.6%, 36.8%)	0.5% (0.4%, 0.6%)	0.7% (0.6%, 0.9%)	0.3% (0.2%, 0.4%)

[a]The denominator is the "*Twitter full archive*" dataset;
[b]The denominator is the combination of "*Twitter search*", "*Twitter filtered stream*", and "*Twitter sampled stream*" datasets.

Table 6 shows the overlapping proportions between each pair of the datasets in Group 2. As shown in **Table 6**, "*Twitter multi-crawler search*" covers a higher proportion than "*Twitter COVID-19 stream*". The overlapping between "*Twitter multi-crawler search*" and "*Twitter COVID-19 stream*" is over 50% at all time scales. When comparing only the tweets containing the top 10 keywords, the overlapping between these two datasets increased to almost 60%. Another interesting finding is that the "*Twitter multi-crawler search*" covered 98.4% of the tweets with the top 10 keywords, which shows that using multiple Twitter search crawlers and multiple queries can increase the data coverage.

Table 6. Data overlapping among datasets in Group 2.						
Scale	**Twitter COVID-19 stream**	**Twitter multi-crawler search**	**Twitter COVID-19 stream vs. Twitter multi-crawler search**	**Twitter COVID-19 stream**	**Twitter multi-crawler search**	**Twitter COVID-19 stream vs. Twitter multi-crawler search**
Experiment	**Comparison of all tweets**			**Comparison of tweets with the top 10 tweets keywords**		
Minute[a]	62.8% (61.9%, 63.8%)	76.0% (74.2%, 77.8%)	51.8% (50.4%, 53.2%)	78.0% (74.9%, 81.1%)	98.4% (97.7%, 99.1%)	77.3% (74.2%, 80.4%)
Minute[b]	73.7% (72.8%,74.6%)	78.9% (77.4%, 80.5%)	52.6% (58.8%, 61.8%)	75.8% (74.0%, 77.6%)	84.5% (83.8%, 85.2%)	60.3% (58.8%, 61.8%)
Hour[b]	67.1% (51.3%, 82.7%)	82.3% (68.5%, 95.5%)	50.5% (30.1%, 62.9%)	74.52% (73.5%, 75.5%)	84.1% (83.7%, 84.5%)	58.6% (57.8%, 59.4%)
Day[b]	69.2% (57.3%, 80.8%)	83.1% (73.4%, 92.6%)	52.1% (43.4, 60.6%)	74.42% (74.2%, 74.7%)	84.5% (84.3%, 84.9%)	58.9% (58.2%, 59.7%)
Week[b]	68.3% (57.3%, 80.8%)	83.3% (69.3%, 96.7%)	50.3% (42.0%, 58.0%)	74.5% (74.2%, 74.8%)	84.3% (84.0%, 84.6%)	58.8% (58.4%, 59.2%)
[a]The denominator is the "*Twitter full archive*" dataset. [b]The denominator is the union of "*Twitter COVID-19 stream*" and "*Twitter multi-crawler search*" datasets.						

Figure 3 visualizes the data overlap in different scenarios across the data collection methods.

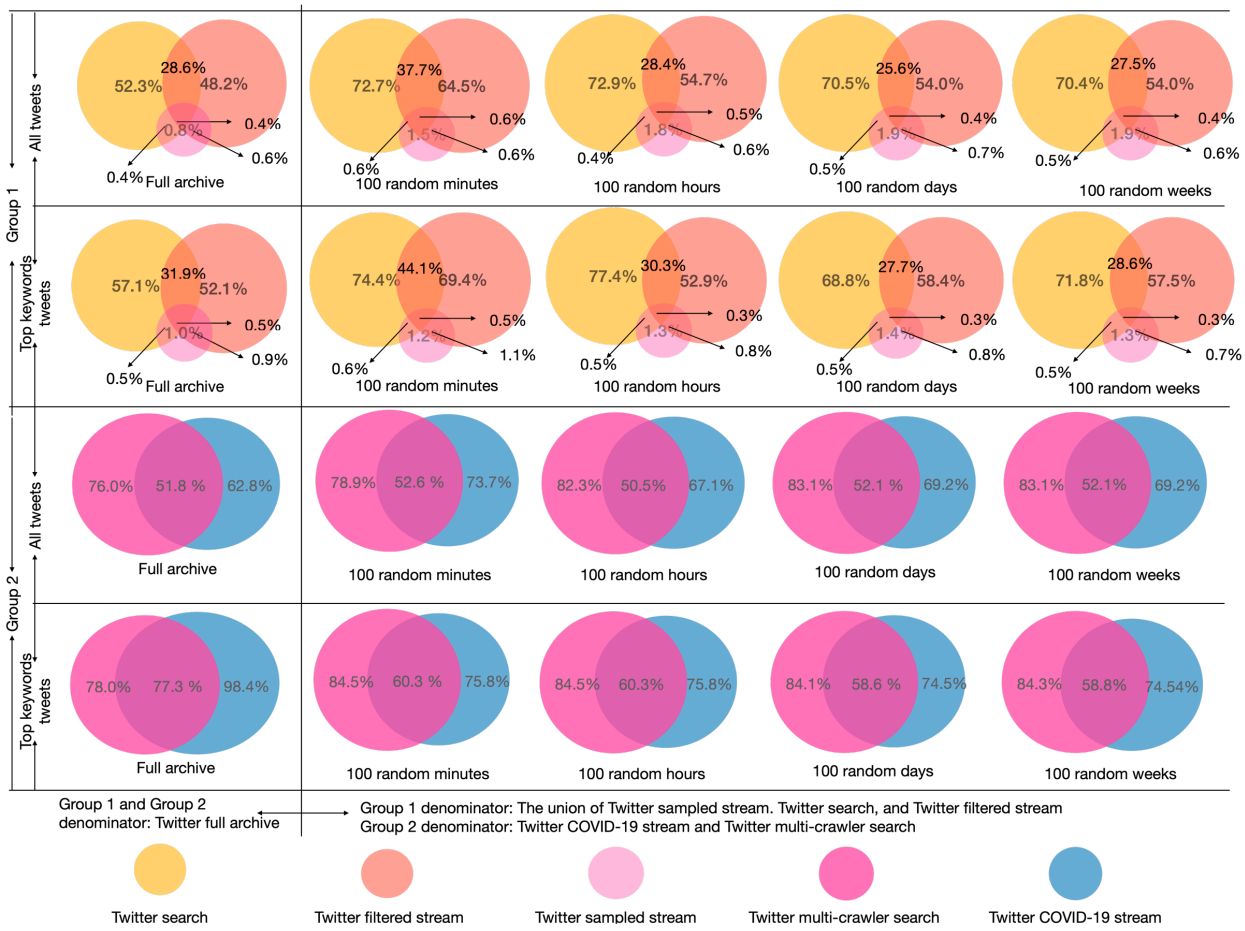

Figure 3. Data overlapping across data collection methods.

RQ2: How representative of the data collected with each of the 5 data collection methods to the gold standard dataset (i.e. from the Twitter full archive)?

We measured the representativeness of each dataset to the "*Twitter full archive*" dataset in terms of the Kendall correlations based on ranking of keywords. As shown in **Table 7**, all the data collection methods in Group 1 have moderate correlations with the "*Twitter full archive*". The two streaming APIs (i.e. "*Twitter sampled stream*" and

"*Twitter filtered stream*") have slightly higher coefficients than "*Twitter search*". In Group 2, "*Twitter COVID-19 stream*" and "*Twitter multi-crawler search*" have very strong correlations with the "*Twitter full archive*" dataset, which coincides by the 98.4% overlapping between "*Twitter multi-crawler search*" and the "*Twitter full archive*".

Table 7. The Kendall correlation coefficients between each dataset to the "*Twitter full archive*" dataset.

Group	Data sources	Coefficient	P value	Correlation[22]
Group 1	Twitter sampled stream vs. Twitter full archive	$\tau = 0.73$	< 0.001	Moderate
	Twitter search vs. Twitter full archive	$\tau = 0.70$	< 0.001	Moderate
	Twitter filtered stream vs. Twitter full archive	$\tau = 0.74$	< 0.001	Moderate
Group 2	Twitter COVID-19 stream vs. Twitter full archive	$\tau = 0.89$	< 0.001	Very strong
	Twitter multi-crawler search vs. Twitter full archive	$\tau = 0.89$	< 0.001	Very strong

Machine learning model sensitivity to data collection bias

RQ3: How does data collection bias (i.e., models trained on different subsets) impact models' performance when applied on other subsets?

Training samples were randomly selected from eight data points as shown in **Figure 2 (B)**. From the 4,800 training samples, our manual annotation classified 113 tweets as irrelevant, 2,064 tweets as promotional information, and 2,623 tweets as consumers' discussions. We trained each our model with a sample dataset and tested model performance on the others datasets. Because of space constraints and CNN models consistently outperformed RF and GBT models, **Table 8** only shows the performance matrix according to CNN models. The training samples were named based the number of the sample points (e.g., the sample from "*Twitter search*" was labelled as "*Sample 2*"), as shown in **Figure 2 (B)**. The performance of each model was reported as a mean F-1 score with associated confidence intervals across 10 runs. As shown in **Table 8,** for CNN models, when we trained the model using a sample from a single dataset of one group and applied the model on samples from the other group, the model performance dropped dramatically, which suggest ML models are sensitive to the time of the data collection. As a comparison, when we trained the CNN model using sample from the "*Twitter full archive*" dataset, the model (i.e., "*Sample 8*") achieved the best average performance and relatively lower variance on other datasets among the eight models. For RF and GBT models, a model can only perform well on the dataset that it trained with, which suggests traditional ML models such as RF and GBT are easily to be overfit on Twitter text classification tasks.

Table 8. Model performance of CNN models trained on different data samples.

CNN Models	Sample 1 (0: 302, 1: 279)[a]	Sample 2 (0: 269, 1: 325)[a]	Sample 3 (0: 301, 1: 283)[a]	Sample 4 (0: 288, 1: 308)[a]	Sample 5 (0: 160, 1: 428)[a]	Sample 6 (0: 216, 1: 364)[a]	Sample 7 (0: 308, 1: 279)[a]	Sample 8 (0: 220, 1: 357)[a]	Mean F1 (SD)
	Group 1				Group 2			Full Archive	
CNN on Sample 1	**0.854 (0.833 ,0.874)**	0.777 (0.757 ,0.797)	0.730 (0.702 ,0.758)	0.792 (0.769 ,0.816)	0.610 (0.574 ,0.647)	0.697 (0.66 ,0.734)	0.689 (0.646 ,0.732)	0.745 (0.707 ,0.782)	0.736 (0.076)
CNN on Sample 2	0.795 (0.77 ,0.821)	0.785 (0.772 ,0.797)	0.688 (0.674, 0.702)	**0.836 (0.817 ,0.854)**	0.674 (0.654, 0.693)	0.762 (0.743 ,0.781)	0.76 (0.74 ,0.78)	0.797 (0.783 ,0.811)	0.762 (0.055)
CNN on Sample 3	0.749 (0.728 ,0.77)	0.765 (0.748 ,0.781)	**0.855 (0.837 ,0.873)**	0.792 (0.782 ,0.802)	0.678 (0.662 ,0.694)	0.785 (0.752, 0.819)	0.745 (0.735 ,0.756)	0.794 (0.786 ,0.802)	0.770 (0.051)
CNN on Sample 4	0.666 (0.645 ,0.687)	0.733 (0.725 ,0.741)	0.667 (0.647 ,0.687)	0.843 (0.836 ,0.849)	0.834 (0.831 ,0.838)	**0.879 (0.874 ,0.883)**	0.816 (0.81 ,0.821)	0.863 (0.858 ,0.868)	0.787 (0.083)
CNN on Sample 5	0.623 (0.608 ,0.639)	0.691 (0.67 ,0.713)	0.604 (0.593 ,0.616)	0.776 (0.755 ,0.797)	0.846 (0.836 ,0.855)	**0.853 (0.837 ,0.87)**	0.808 (0.791 ,0.826)	0.852 (0.836 ,0.869)	0.756 (0.100)
CNN on Sample 6	0.605 (0.591 ,0.62)	0.713 (0.7 ,0.725)	0.602 (0.59 ,0.613)	0.786 (0.771 ,0.801)	**0.858 (0.853 ,0.864)**	0.846 (0.832 ,0.86)	0.817 (0.811 ,0.823)	0.857 (0.849 ,0.864)	0.760 (0.100)
CNN on Sample 7	0.681 (0.661 ,0.701)	0.743 (0.735 ,0.752)	0.708 (0.694 ,0.721)	0.815 (0.808 ,0.823)	0.811 (0.798 ,0.825)	0.850 (0.837 ,0.862)	0.833 (0.824 ,0.841)	**0.852 (0.845 ,0.86)**	0.786 (0.065)
CNN on Sample 8	0.704 (0.683 ,0.724)	0.761 (0.74 ,0.781)	0.704 (0.68 ,0.728)	0.824 (0.806 ,0.841)	0.835 (0.825 ,0.844)	0.853 (0.848 ,0.859)	0.819 (0.814 ,0.824)	**0.870 (0.866 ,0.883)**	**0.796 (0.064)**

[a]Number of tweets for label 0 (i.e. consumers' discussions) and 1 (i.e. promotional information).

Discussion and conclusion

In this study, we aimed to assess data collection bias among different data collection methods provided by Twitter, identify the representativeness of each data collection methods compared with the "*Twitter full archive*", and test ML models' sensitivities to data collection bias, through answering three research questions (RQs).

For **RQ1**, we found that, first, the "*Twitter multi-crawler search*" can effectively collect more tweets than a single "*Twitter search*" crawler, even with the same set of keywords, in terms of both data volume and overlapping

proportions (i.e. covering more samples of the Twitter universe), suggesting that (1) Twitter's internal subsampling strategies might not be consistent across different endpoints due to rate limits, and (2) using multiple crawler and multiple queries is a way to work around the API rate limits.[16] Nevertheless, using multiple crawlers requires researchers to have multiple Twitter accounts, and as Twitter has strengthened its identity verification process, especially for developer accounts (e.g., each phone number can only register a single Twitter account), this presents a challenge. We also found that overlapping proportions between the "*Twitter filtered stream*" vs. "*Twitter search*" in Group 1 and the "*Twitter COVID-19 stream*" vs. "*Twitter multi-crawler search*" in Group 2 increased when we compared only the tweets that included the top 10 keywords, leading to our RQ2, the representativeness of the tweets collected by different data collection methods.

To answer **RQ2,** we assessed the Kendall correlations between the "*Twitter full archive*" benchmark dataset and each of the other five datasets. Among the five other datasets, the "*Twitter filtered stream*", "*Twitter sampled stream*", and "*Twitter search*" datasets all have moderate correlations, which indicates that even though these crawlers only collect subsamples of the tweets in Twitter data warehouse, these subsamples are still representative and can be used to "*identify and track trends, monitor general sentiment, monitor global events, and much more*" as claimed in a Twitter API document.[13] Somewhat surprisingly, the "*Twitter multi-crawler search*" ($\tau = 0.89$) has the same level of correlation to "*Twitter full archive*" as the "*Twitter COVID-19 stream*" ($\tau = 0.89$), which shows the effectiveness of the multi-crawlers and multi-queries strategy. The "*Twitter COVID-19 stream*" was designed based on Twitter's internal COVID-19 Tweet annotation and parameters, which they "*believe deliver a comprehensive view of the conversation around this topic.*" The effectiveness of the "*Twitter multi-crawler search*" gives us some level of confidence that studies that use the "*Twitter multi-crawler search"* for other public health studies, where an endpoint like "*Twitter COVID-19 stream*" does not exist, can still be conducted.

To answer **RQ3**, we randomly selected eight training samples for building ML models from different parts of the datasets as shown in **Figure 2 (B)**. Among the eight training samples, seven samples were selected either from a single dataset or from overlapping portions within each group, and one sample was selected from the "*Twitter full archive*" dataset covering the time ranges of both Group 1 and Group 2. We found that (1) in general, models trained on one dataset cannot perform well on the samples from the other group (i.e., time shifts have a significant impact on model performance); (2) CNN models can achieve the highest average performance with relatively reliable consistence when using sample from the "*Twitter full archive*". These results suggest that it is necessary to selection training samples from a representative dataset, and as the time progress, it is important to retrain of ML models with new datasets; and (3) traditional ML models such as RF and GBT can only achieve reasonable performance on the samples that they trained with. This indicates that compared with CNN models, RF and GBT are more easily to be overfitted. Thus, when conducting social media analyses, more thorough experimentation and testing of the selected models and their underlying assumptions of the data is necessary.

We also recognized the limitations of our study. First, we cannot recover the "*Twitter full archive*" dataset back to the time when we collected the other datasets, because of issues such as deleted tweets and suspended accounts. Thus, our measures of overlapping between other datasets to the full archive are only approximates. Second, many other factors that may affect the Twitter data collections such as the number of crawlers running on a single machine (i.e., competing of CPU cycles), and the reliability of the Internet connections among others. Third, many other factors may affect the ML model performance as well such as the sample size of the training samples, data preprocessing methods, hyper-parameter tuning, and data imbalance issues. Fourth, there are many other types of biases in Twitter studies, such as demographic bias and keyword bias.[23] Weeg *et al.* (2015) mitigated demographic bias of Twitter data by stratifying Twitter users based on geographic distributions.[24] Kim *et al.* (2016) measured the quality of data collection in two aspects: 1) retrieval precision (i.e., "*precision measures how much of the retrieved data is not garbage*" and 2) retrieval recall (i.e., "*recall measures how much of the relevant data is retrieved*") and proposed a conceptual framework for the filtering and quality evaluation of social data.[25] Those biases and potential methods are worth investigating in future research.

In conclusion, our study assessed the data collection bias, evaluated the representativeness of multiple data collection methods, and tested ML models' sensitivity to data collection bias. Data and model bias issues are often ignored in social media studies. However, to really use social media such as Twitter as a reliable data source for future research, we must find ways to address (or at least assess) data and model biases.

Acknowledgment

This work was supported in part by NSF Award #1734134.

References

1. Times TNY. Coronavirus in the U.S.: Latest Map and Case Count. The New York Times [Internet]. 2020 Jul 20 [cited 2021 Mar 9]; Available from: https://www.nytimes.com/interactive/2020/us/coronavirus-us-cases.html
2. CDC. COVID-19 and Your Health [Internet]. Centers for Disease Control and Prevention. 2021 [cited 2021 Mar 9]. Available from: https://www.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/prevention.html
3. ALYSSA NEWCOMB. Conferences go online amid coronavirus fears—minus the hallway schmoozing. Fortune [Internet]. 2020 Mar 10 [cited 2021 Mar 9]; Available from: https://fortune.com/2020/03/10/zoom-sxsw-coronavirus/
4. WHO Is Fighting False COVID Info On Social Media. How's That Going? [Internet]. NPR.org. [cited 2021 Mar 9]. Available from: https://www.npr.org/sections/goatsandsoda/2021/02/09/963973675/who-is-fighting-false-covid-info-on-social-media-hows-that-going
5. Kouzy R, Abi Jaoude J, Kraitem A, El Alam MB, Karam B, Adib E, et al. Coronavirus Goes Viral: Quantifying the COVID-19 Misinformation Epidemic on Twitter. Cureus [Internet]. 2020 Mar 13 [cited 2021 Mar 9]; Available from: https://www.cureus.com/articles/28976-coronavirus-goes-viral-quantifying-the-covid-19-misinformation-epidemic-on-twitter
6. Xue J, Chen J, Hu R, Chen C, Zheng C, Su Y, et al. Twitter Discussions and Emotions About the COVID-19 Pandemic: Machine Learning Approach. J Med Internet Res. 2020 Nov 25;22(11):e20550.
7. Mackey T, Purushothaman V, Li J, Shah N, Nali M, Bardier C, et al. Machine Learning to Detect Self-Reporting of Symptoms, Testing Access, and Recovery Associated With COVID-19 on Twitter: Retrospective Big Data Infoveillance Study. JMIR Public Health Surveill. 2020 Jun 8;6(2):e19509.
8. Pfeffer J, Mayer K, Morstatter F. Tampering with Twitter's Sample API. EPJ Data Sci. 2018 Dec;7(1):1–21.
9. Jun I, Zhao Y, He X, Gollakner R, Court C, Munoz O, et al. Understanding Perceptions and Attitudes toward Genetically Modified Organisms on Twitter. In: International Conference on Social Media and Society [Internet]. Toronto ON Canada: ACM; 2020 [cited 2021 Mar 9]. p. 291–8. Available from: https://dl.acm.org/doi/10.1145/3400806.3400839
10. Bian J, Zhao Y, Salloum RG, Guo Y, Wang M, Prosperi M, et al. Using Social Media Data to Understand the Impact of Promotional Information on Laypeople's Discussions: A Case Study of Lynch Syndrome. J Med Internet Res. 2017 Dec 13;19(12):e414.
11. Zhang H, Wheldon C, Dunn AG, Tao C, Huo J, Zhang R, et al. Mining Twitter to assess the determinants of health behavior toward human papillomavirus vaccination in the United States. J Am Med Inform Assoc JAMIA. 2020 Feb 1;27(2):225–35.
12. Twitter. Twitter Search API [Internet]. 2021 [cited 2021 Mar 4]. Available from: https://developer.twitter.com/en/docs/twitter-api/v1/tweets/search/api-reference/get-search-tweets
13. Twitter. Sampled stream | Twitter API [Internet]. 2021 [cited 2021 Mar 10]. Available from: https://developer.twitter.com/en/docs/twitter-api/tweets/sampled-stream/introduction
14. Twitter. Twitter Filtered Stream API [Internet]. 2021 [cited 2021 Mar 4]. Available from: https://developer.twitter.com/en/docs/twitter-api/v1/tweets/filter-realtime/guides/connecting
15. Twitter. COVID-19 stream [Internet]. 2021 [cited 2021 Mar 4]. Available from: https://developer.twitter.com/en/docs/labs/covid19-stream/overview
16. Twitter. Rate limits: Standard v1.1 [Internet]. 2021 [cited 2021 Mar 5]. Available from: https://developer.twitter.com/en/docs/twitter-api/v1/rate-limits
17. Twitter. Full-archive search quick start [Internet]. 2021 [cited 2021 Mar 4]. Available from: https://developer.twitter.com/en/docs/twitter-api/tweets/search/quick-start/full-archive-search
18. Pennington J, Socher R, Manning CD. GloVe: Global Vectors for Word Representation. In: Empirical Methods in Natural Language Processing (EMNLP) [Internet]. 2014. p. 1532–43. Available from: http://www.aclweb.org/anthology/D14-1162
19. Kim Y. Convolutional Neural Networks for Sentence Classification. In Association for Computational Linguistics; 2014 [cited 2017 Sep 21]. p. 1746–51. Available from: http://aclweb.org/anthology/D14-1181
20. Liaw A, Wiener M. Classification and Regression by randomForest. R News. 2002;2(3):18–22.
21. Friedman JH. Greedy function approximation: a gradient boosting machine. Ann Stat. 2001;1189–232.
22. Akoglu H. User's guide to correlation coefficients. Turk J Emerg Med. 2018 Aug 7;18(3):91–3.
23. Jiang Y, Li Z, Ye X. Understanding demographic and socioeconomic biases of geotagged Twitter users at the county level. Cartogr Geogr Inf Sci. 2019 May 4;46(3):228–42.
24. Weeg C, Schwartz HA, Hill S, Merchant RM, Arango C, Ungar L. Using Twitter to Measure Public Discussion of Diseases: A Case Study. JMIR Public Health Surveill. 2015 Jun 26;1(1):e6.
25. Kim Y, Huang J, Emery S. Garbage in, Garbage Out: Data Collection, Quality Assessment and Reporting Standards for Social Media Data Use in Health Research, Infodemiology and Digital Disease Detection. J Med Internet Res. 2016 Feb 26;18(2):e41.