Original Paper

Toward Using Twitter for Tracking COVID-19: A Natural Language Processing Pipeline and Exploratory Data Set

Ari Z Klein, PhD; Arjun Magge, PhD; Karen O'Connor, MS; Jesus Ivan Flores Amaro, BS; Davy Weissenbacher, PhD; Graciela Gonzalez Hernandez, PhD

Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, United States

Corresponding Author:
Ari Z Klein, PhD
Department of Biostatistics, Epidemiology, and Informatics
Perelman School of Medicine
University of Pennsylvania
421A Blockley Hall
423 Guardian Dr
Philadelphia, PA, 19104
United States
Phone: 1 215 746 1101
Email: ariklein@pennmedicine.upenn.edu

Abstract

Background: In the United States, the rapidly evolving COVID-19 outbreak, the shortage of available testing, and the delay of test results present challenges for actively monitoring its spread based on testing alone.

Objective: The objective of this study was to develop, evaluate, and deploy an automatic natural language processing pipeline to collect user-generated Twitter data as a complementary resource for identifying potential cases of COVID-19 in the United States that are not based on testing and, thus, may not have been reported to the Centers for Disease Control and Prevention.

Methods: Beginning January 23, 2020, we collected English tweets from the Twitter Streaming application programming interface that mention keywords related to COVID-19. We applied handwritten regular expressions to identify tweets indicating that the user potentially has been exposed to COVID-19. We automatically filtered out “reported speech” (eg, quotations, news headlines) from the tweets that matched the regular expressions, and two annotators annotated a random sample of 8976 tweets that are geo-tagged or have profile location metadata, distinguishing tweets that self-report potential cases of COVID-19 from those that do not. We used the annotated tweets to train and evaluate deep neural network classifiers based on bidirectional encoder representations from transformers (BERT). Finally, we deployed the automatic pipeline on more than 85 million unlabeled tweets that were continuously collected between March 1 and August 21, 2020.

Results: Interannotator agreement, based on dual annotations for 3644 (41%) of the 8976 tweets, was 0.77 (Cohen κ). A deep neural network classifier, based on a BERT model that was pretrained on tweets related to COVID-19, achieved an F_1-score of 0.76 (precision=0.76, recall=0.76) for detecting tweets that self-report potential cases of COVID-19. Upon deploying our automatic pipeline, we identified 13,714 tweets that self-report potential cases of COVID-19 and have US state–level geolocations.

Conclusions: We have made the 13,714 tweets identified in this study, along with each tweet’s time stamp and US state–level geolocation, publicly available to download. This data set presents the opportunity for future work to assess the utility of Twitter data as a complementary resource for tracking the spread of COVID-19.

(J Med Internet Res 2021;23(1):e25314) doi: 10.2196/25314

KEYWORDS

natural language processing; social media; data mining; COVID-19; coronavirus; pandemics; epidemiology; infodemiology

Introduction

In the United States, the rapidly evolving COVID-19 outbreak, the shortage of available testing, and the delay of test results have presented challenges for actively monitoring the spread of COVID-19 based on testing alone. An approach that has emerged for detecting cases without the need for extensive testing relies on voluntary self-reports of symptoms from the

general population [1]. Considering that nearly one of every four adults in the United States already uses Twitter, and nearly half of them use it on a daily basis [2], researchers have begun exploring tweets for mentions of COVID-19 symptoms [3-8]. However, considering the incubation period of COVID-19 [9], detecting cases based on symptoms may not maximize the potential of Twitter data for real-time monitoring. The objective of this study was to develop, evaluate, and deploy a natural language processing (NLP) pipeline that automatically collects tweets reporting personal information more broadly—that is, beyond symptoms—that might indicate exposure to COVID-19 in the United States. In this paper, we present a publicly available data set containing 13,714 tweets that were identified by our automatic NLP pipeline between March 1 and August 21, 2020, with each tweet's time stamp and US state–level geolocation. This data set presents the opportunity to explore the use of Twitter data as a complementary resource "to understand and model the transmission and trajectory of COVID-19" [10].

Methods

Data Collection and Annotation

The Institutional Review Board (IRB) of the University of Pennsylvania reviewed this study and deemed it to be exempt human subjects research under Category (4) of Paragraph (b) of the US Code of Federal Regulations Title 45 Section 46.101 for publicly available data sources (45 CFR §46.101(b)(4)).

Between January 23 and March 20, 2020, we collected more than 7 million publicly available tweets that mention keywords related to COVID-19, are posted in English, are not retweets, and are geo-tagged or have user profile location metadata. We developed handwritten regular expressions (Multimedia Appendix 1)—search patterns designed to automatically match text strings—to identify a subset of the 7 million tweets that indicate that the user potentially has been exposed to COVID-19. Our query patterns were designed primarily to help identify potential cases of COVID-19 that are not based on testing and, thus, may not have been reported to the Centers for Disease Control and Prevention (CDC) [11]. The regular expressions matched approximately 160,000 (2%) of the 7 million tweets. Approximately 30,000 (19%) of the 160,000 matching tweets were then automatically removed using a system we developed in recent work [12] for filtering out "reported speech" (eg, quotations, news headlines) from health-related social media data.

In preliminary work [13], two annotators annotated a random sample of 10,000 of the 130,000 filtered tweets, and annotation guidelines (Multimedia Appendix 2) were developed to help the annotators distinguish between three classes of tweets. However, since then, we have removed 1024 of the annotated tweets that were collected from the Twitter Streaming application programming interface (API) based on a keyword that we have stopped using, and we have unified two of the classes. "Potential case" tweets include those that indicate that the user or a member of the user's household was denied testing for COVID-19, showing symptoms of COVID-19, potentially exposed to presumptive or confirmed cases of COVID-19, or had had experiences that pose a higher risk of exposure to COVID-19. "Other" tweets are related to COVID-19 and may discuss topics such as testing, symptoms, traveling, or social distancing, but do not indicate that the user or a member of the user's household may be infected. Among the 8976 tweets, 3644 (41%) were annotated by both annotators. Upon resolving the annotators' disagreements, 1456 (16%) of the tweets were annotated as "potential case" and 7520 (84%) as "other." Textbox 1 presents (slightly modified) sample tweets that match our handwritten regular expressions and were manually annotated as "potential case."

Textbox 1. Sample (slightly modified) tweets that match our handwritten regular expressions and were manually annotated as potential cases of COVID-19.

1. Nearly two weeks ago I had a fever, sore throat, runny nose, and cough. I want to know if it was coronavirus or just the common cold
2. My coworker in next office probably has #coronavirus. He and his wife have the symptoms, but they went to the hospital to get tested and were refused.
3. This girl in my class had the coronavirus, so I'm making an appointment with my doctor for a check up
4. Pretty sure I had a patient tonight with Coronavirus. Had all the symptoms and tested negative for the flu.
5. Why can celebrities, sports athletes & politicians without symptoms get tested, but my symptomatic child who has a compromised immune system cannot? #coronavirus
6. Since getting back from Seattle I've been sick and want to get a #coronavirus check. Called my PCP, they said to call health dept. Called them, they said I need to go thru my PCP. Called my PCP again, they said they can't help me
7. I'm convinced I have coronavirus. I've been to NYC, Phoenix, and San Diego in the last few weeks. I have a cough, a runny nose, and I'm really hot #covid19
8. Scared of the coronavirus because I have a sore throat and a headache I think its just a cold but I take the tube 4 times a day
9. Can't even get testing SCHEDULED while self-quarantined (my decision) and having coronavirus symptoms I take train thru New Rochelle to Manhattan
10. I have a bad cold. I went to the doctor, got some medications, the norm. But they couldn't rule out coronavirus because they don't have the tests.

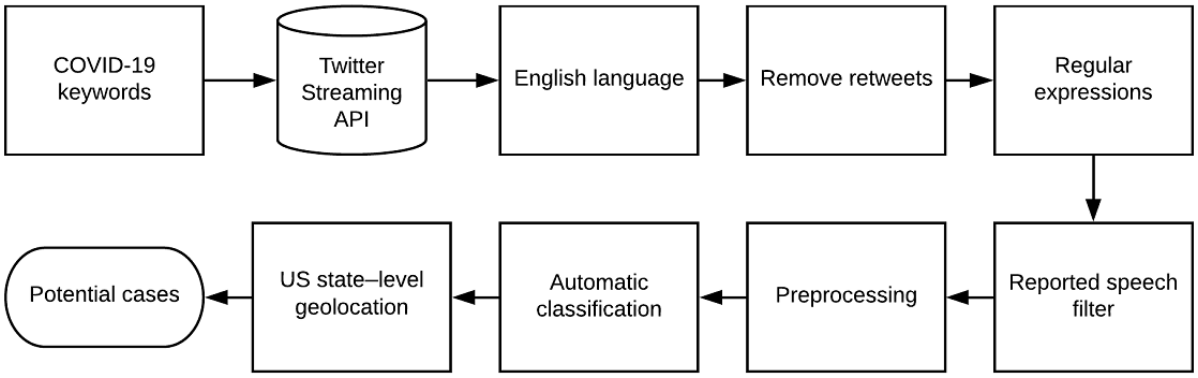
As Textbox 1 illustrates, our handwritten regular expressions are based on query patterns designed to identify tweets that report personal information that may be useful for tracking potential cases of COVID-19, including not only symptoms

(tweet 1), but also exposure to potential cases and a lack of access to COVID-19 testing. For example, our regular expressions retrieve tweets reporting that the user may have come in contact with coworkers (tweet 2), classmates (tweet 3), patients (tweet 4), and family members (tweet 5) who may have COVID-19, and potential exposure to COVID-19 through traveling (tweets 6 and 7) and commutes (tweets 8 and 9). Our regular expressions also retrieve tweets reporting that the user (tweet 9 and 10), a family member (tweet 5), or someone else that the user has been in contact with (tweet 2) was denied access to testing, even though they are sick. Since none of the tweets in Textbox 1 report being tested for or diagnosed with COVID-19, they represent potential cases that may not have been reported to the CDC.

Automatic Classification and Geolocation

We split the 8976 annotated tweets into 80% (7181 tweets) and 20% (1795 tweets) random sets—a training set (Multimedia Appendix 3) and held-out test set, respectively—for automatic classification. We used the *ktrain* [14] Python library to train and evaluate two supervised deep neural network classifiers based on bidirectional encoder representations from transformers (BERT): BERT-Base-Uncased [15] and COVID-Twitter-BERT [16]. After feeding the sequence of tweet tokens to BERT, the encoded representation is passed to a dropout layer (dropping rate of 0.1), followed by a dense layer with 2 units and a softmax activation, which predicts the class for each tweet. For training, we used Adam optimization with rate decay and warm-up. We used a batch size of 64, training runs for 3 epochs, and a maximum learning rate of 1×10^{-5}. We fine-tuned all layers of the transformer model with our annotated tweets. Prior to automatic classification, we preprocessed the tweets by normalizing usernames and URLs, and lowercasing the text. Figure 1 illustrates our automatic NLP pipeline for detecting tweets that indicate potential cases of COVID-19 in the United States. We deployed the pipeline on more than 85 million unlabeled tweets that were continuously collected between March 1 and August 21, 2020. We used Carmen [17] to infer the geolocation—at the US state level—of tweets that the classifier predicted as potential cases.

Figure 1. Automatic natural language processing (NLP) pipeline for detecting tweets that self-report potential cases of COVID-19 in the United States.



Results

Interannotator agreement, based on dual annotations for 3644 (41%) of the 8976 tweets, was 0.77 (Cohen κ), considered "substantial agreement" [18]. We evaluated two deep neural network classifiers on a held-out test set of 1795 (20%) of the 8976 tweets. The classifier based on the BERT-Base-Uncased pretrained model achieved an F_1-score of 0.70 (precision=0.72, recall=0.67) for the "potential case" class, and the classifier based on the COVID-Twitter-BERT pretrained model achieved an F_1-score of 0.76 (precision=0.76, recall=0.76), where:

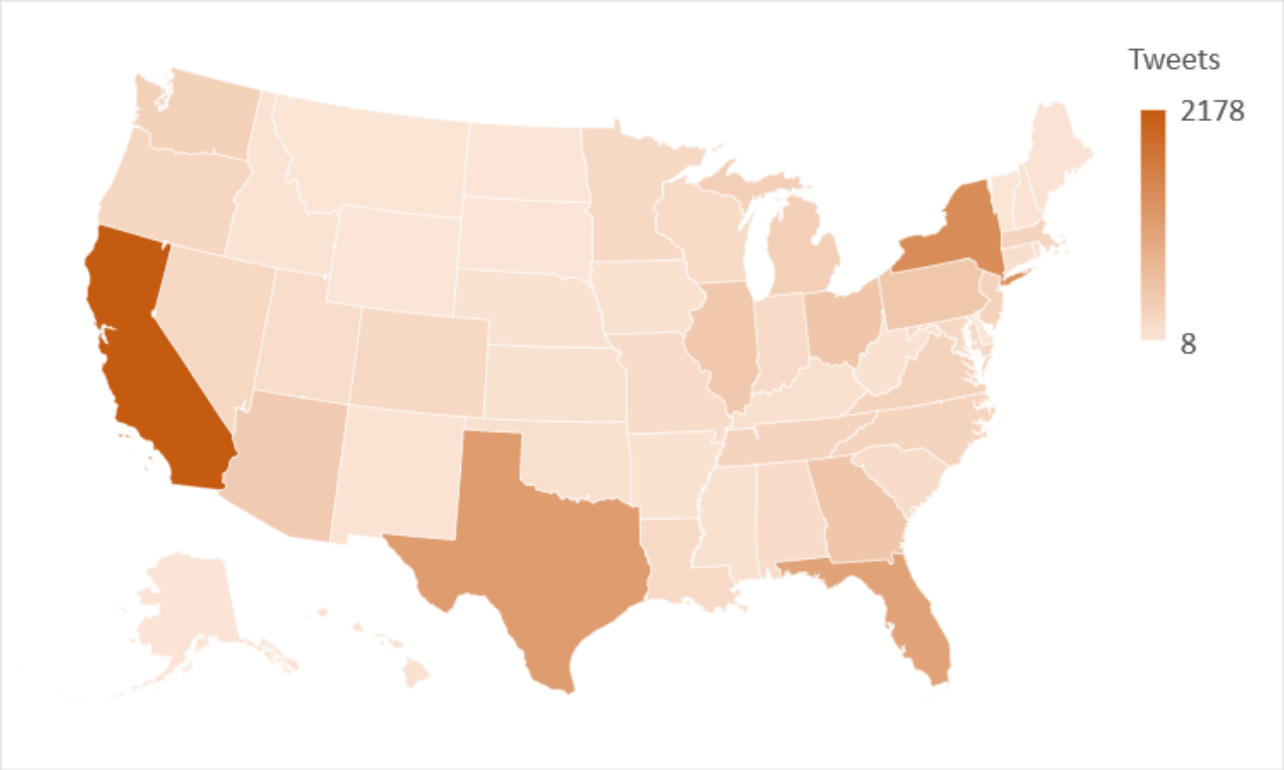
$$F_1\text{-score} = \frac{2\ x\ recall\ x\ precision}{recall + precision}$$

$$Precision = \frac{true\ positives}{true\ positives + false\ positives}$$

$$Recall = \frac{true\ positives}{true\ positives + false\ negatives}$$

We deployed our automatic pipeline, using the COVID-Twitter-BERT classifier, on more than 85 million unlabeled tweets that were continuously collected from the Twitter Streaming API between March 1 and August 21, 2020. Among the subset of tweets that were posted in English, not retweets, matched the regular expressions, and were not filtered out as reported speech, the COVID-Twitter-BERT classifier detected 13,714 "potential case" tweets for which Carmen inferred a US state–level geolocation. Figure 2 illustrates the ranges of "potential case" tweets that were automatically detected per state. We automatically detected "potential case" tweets from all 50 states, with the highest numbers posted in California, New York, Texas, and Florida.

Figure 2. Tweets self-reporting potential cases of COVID-19 in the United States, by state, between March 1 and August 21, 2020.



Discussion

Principal Findings

While Twitter data has been used to identify self-reports of symptoms by people who have tested positive for COVID-19 [3,4], the shortage of available testing and the delay of test results in the United States motivated us to assess whether Twitter data could be scaled to identify potential cases of COVID-19 that are not based on testing and, thus, may not have been reported to the CDC. There are studies that have not limited their exploration of COVID-19 symptoms on Twitter to users who have tested positive for COVID-19 [5-8]; however, limiting the detection of potential cases to symptoms may still underutilize the information available on Twitter. Our automatic NLP pipeline has detected potential cases of COVID-19 across the entire United States that are neither based on testing nor limited to symptoms, providing the opportunity to explore the utility of Twitter data more broadly as a complementary resource for tracking the spread of COVID-19. An analysis based on this data set is beyond the scope of this study. The 13,714 "potential case" tweets identified in this study can be downloaded using a Python script [19] and the input file in Multimedia Appendix 4, which contains the user ID, tweet ID, time stamp, and inferred state-level geolocation for each tweet. The script downloads the tweets that are still publicly available.

Conclusions

This paper presented an automatic NLP pipeline that was used to identify 13,714 tweets self-reporting potential cases of COVID-19 in the United States between March 1 and August 21, 2020, that may not have been reported to the CDC. This publicly available data set presents the opportunity for future work to assess the utility of Twitter data as a complementary resource for tracking the spread of COVID-19.

Acknowledgments

AZK contributed to the methodology, formal analysis, investigation, data curation, and writing the original draft. AM contributed to the software development, formal analysis, investigation, and writing the original draft. KO contributed to the data curation and writing (review and editing). JIFA contributed to the software development and writing (review and editing). DW contributed to the software development, formal analysis, investigation, and writing (review and editing). GGH contributed to the conceptualization, writing (review and editing), supervision, and funding acquisition. The authors would like to thank Alexis Upshur for contributing to annotating the Twitter data. This work was supported by the National Institutes of Health (NIH) National Library of Medicine (NLM; grant number R01LM011176) and National Institute of Allergy and Infectious Diseases (NIAID; grant number R01AI117011).

Conflicts of Interest

None declared.

Multimedia Appendix 1

Regular expressions.
[TXT File , 3 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Annotation guidelines.
[PDF File (Adobe PDF File), 1060 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Training data.
[TXT File , 249 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Exploratory Twitter data set.
[TXT File , 851 KB-Multimedia Appendix 4]

References

1. Menni C, Valdes AM, Freidin MB, Sudre CH, Nguyen LH, Drew DA, et al. Real-time tracking of self-reported symptoms to predict potential COVID-19. Nat Med 2020 Jul 11;26(7):1037-1040 [FREE Full text] [doi: 10.1038/s41591-020-0916-2] [Medline: 32393804]
2. Smith A, Anderson M. Social media use in 2018. Pew Research Center. 2018 Mar 01. URL: https://www.pewresearch.org/internet/2018/03/01/social-media-use-in-2018/ [accessed 2020-09-29]
3. Sarker A, Lakamana S, Hogg-Bremer W, Xie A, Al-Garadi M, Yang Y. Self-reported COVID-19 symptoms on Twitter: an analysis and a research resource. J Am Med Inform Assoc 2020 Aug 01;27(8):1310-1315. [doi: 10.1093/jamia/ocaa116] [Medline: 32620975]
4. Jeon J, Baruah G, Sarabadani S, Palanica A. Identification of Risk Factors and Symptoms of COVID-19: Analysis of Biomedical Literature and Social Media Data. J Med Internet Res 2020 Oct 02;22(10):e20509 [FREE Full text] [doi: 10.2196/20509] [Medline: 32936770]
5. 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 08;6(2):e19509 [FREE Full text] [doi: 10.2196/19509] [Medline: 32490846]
6. Panuganti BA, Jafari A, MacDonald B, DeConde AS. Predicting COVID-19 Incidence Using Anosmia and Other COVID-19 Symptomatology: Preliminary Analysis Using Google and Twitter. Otolaryngol Head Neck Surg 2020 Sep;163(3):491-497 [FREE Full text] [doi: 10.1177/0194599820932128] [Medline: 32484425]
7. Guntuku SC, Sherman G, Stokes DC, Agarwal AK, Seltzer E, Merchant RM, et al. Tracking Mental Health and Symptom Mentions on Twitter During COVID-19. J Gen Intern Med 2020 Sep 07;35(9):2798-2800 [FREE Full text] [doi: 10.1007/s11606-020-05988-8] [Medline: 32638321]
8. Guo J, Radloff CL, Wawrzynski SE, Cloyes KG. Mining twitter to explore the emergence of COVID-19 symptoms. Public Health Nurs 2020 Nov 16;37(6):934-940. [doi: 10.1111/phn.12809] [Medline: 32937679]
9. Lauer SA, Grantz KH, Bi Q, Jones FK, Zheng Q, Meredith HR, et al. The Incubation Period of Coronavirus Disease 2019 (COVID-19) From Publicly Reported Confirmed Cases: Estimation and Application. Annals of Internal Medicine 2020 May 05;172(9):577-582. [doi: 10.7326/m20-0504]
10. Merchant RM, Lurie N. Social Media and Emergency Preparedness in Response to Novel Coronavirus. JAMA 2020 May 26;323(20):2011-2012. [doi: 10.1001/jama.2020.4469] [Medline: 32202611]
11. United States COVID-19 cases and deaths by state. Centers for Disease Control and Prevention. URL: https://covid.cdc.gov/covid-data-tracker/#cases_totalcases [accessed 2020-09-29]
12. Klein AZ, Cai H, Weissenbacher D, Levine LD, Gonzalez-Hernandez G. A natural language processing pipeline to advance the use of Twitter data for digital epidemiology of adverse pregnancy outcomes. Journal of Biomedical Informatics: X 2020 Dec;8:100076. [doi: 10.1016/j.yjbinx.2020.100076]
13. Klein A, Magge A, O'Connor K, Cai H, Weissenbacher D, Gonzalez-Hernandez G. A chronological and geographical analysis of personal reports of COVID-19 on Twitter. medRxiv Preprint published online on April 24, 2020. [doi: 10.1101/2020.04.19.20069948]
14. Maiya AS. ktrain: a low-code library for augmented machine learning. arXiv Preprint posted online on April 19, 2020. [FREE Full text]
15. Devlin J, Cheng M, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. 2019 Presented at: 17th Annual Conference of the North American Chapter of the Association for Computational Linguistics:

Human Language Technologies (NAACL-HLT); June 2-7, 2019; Minneapolis, MN p. 4171-4186. [doi: 10.18653/v1/N19-1423]
16. Müller M, Salathé M, Kummervold P. COVID-Twitter-BERT: a natural language processing model to analyse COVID-19 content on Twitter. arXiv Preprint posted online on May 15, 2020. [FREE Full text]
17. Drezde M, Paul M, Bergsma S, Tran H. Carmen: a Twitter geo-location system with applications to public health. 2013 Presented at: Association for the Advancement of Artificial Intelligence (AIII) 2013 Workshop Expanding the Boundaries of Health Informatics Using AI; July 14-15, 2013; Bellevue, WA, USA.
18. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. Fam Med 2005 May;37(5):360-363 [FREE Full text] [Medline: 15883903]
19. Study data set. Bitbucket. URL: https://bitbucket.org/pennhlp/twitter_data_download/src/master/ [accessed 2021-01-18]

Abbreviations

API: application programming interface
BERT: bidirectional encoder representations from transformers
CDC: Centers for Disease Control and Prevention
NLP: natural language processing

Edited by G Eysenbach; submitted 27.10.20; peer-reviewed by K Verspoor, V Foufi, X Ji, L Sheets; comments to author 05.12.20; revised version received 14.12.20; accepted 14.12.20; published 22.01.21

Please cite as:
Klein AZ, Magge A, O'Connor K, Flores Amaro JI, Weissenbacher D, Gonzalez Hernandez G
Toward Using Twitter for Tracking COVID-19: A Natural Language Processing Pipeline and Exploratory Data Set
J Med Internet Res 2021;23(1):e25314
URL: http://www.jmir.org/2021/1/e25314/
doi: 10.2196/25314
PMID: 33449904