Toward Using Twitter Data for Tracking COVID-19 in the United States

Ari Z. Klein, Arjun Magge, Karen O'Connor, Ivan Flores, Davy Weissenbacher, Graciela Gonzalez-Hernandez
Department of Biostatistics, Epidemiology & Informatics, University of Pennsylvania
Health Language Processing Center (https://healthlanguageprocessing.org)
{ariklein, gragon}@pennmedicine.upenn.edu

Background

- 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].
- Tweets have been explored for mentions of COVID-19 symptoms; however, considering the incubation period of COVID-19[2], detecting cases based on symptoms may not maximize the potential of Twitter data for real-time monitoring.

Objectives

- To develop, evaluate, and deploy an automatic natural language processing (NLP) pipeline that collects tweets reporting personal information more broadly (i.e., beyond symptoms) that might indicate exposure to COVID-19 in the United States.

Conclusions

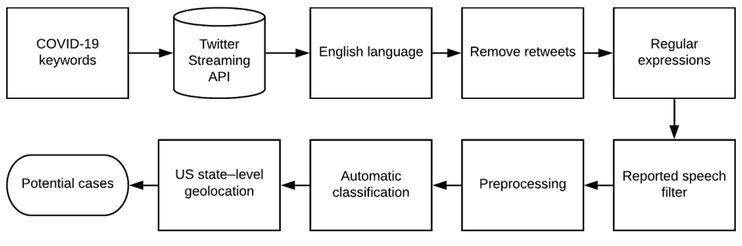
- Our pipeline was used to identify 13,714 tweets self-reporting potential cases of COVID-19 in the United States that may not have been reported to the CDC. This publicly available data set presents the opportunity for future work to explore the utility of Twitter data as a complementary resource for tracking the spread of COVID-19.

Methods

- *Data collection.* We collected tweets from the Twitter Streaming API that mention keywords related to COVID-19. We developed handwritten regular expressions to identify a subset of the tweets indicating that the user potentially has been exposed to COVID-19—in particular, potential cases that are not based on testing and, thus, may not have reported to the CDC.
- *Annotation*. Two annotators manually distinguished "potential case" (examples below) and "other" tweets in a random sample of 8976 of the tweets that matched the regular expressions, were posted in English, were not retweets, and were not "reported speech"[3] (e.g., news headlines), with inter-annotator agreement (Cohen's kappa) of 0.77.

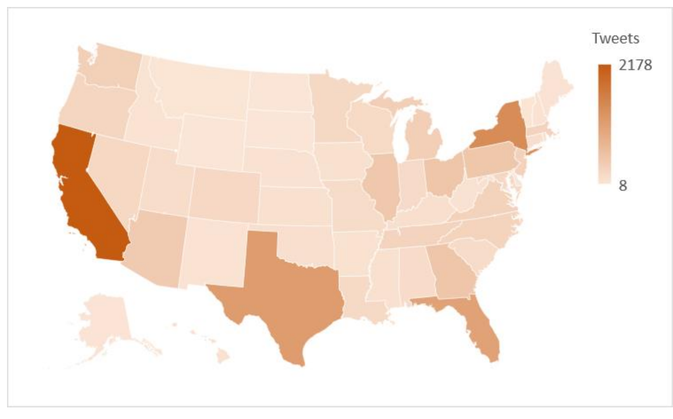
I'm convinced I have coronavirus. I've been to NYC, Phoenix, and San Diego in the last few weeks. I have a cough, a runny nose, and I'm really hot #covid19
I have a bad cold. I went to the doctor, got some medications, the norm. But they couldn't rule out coronavirus because they don't have the tests.
This girl in my class had the coronavirus, so I'm making an appointment with my doctor for a check up

- *Automatic classification and geolocation.* We used the annotated tweets to train and evaluate deep neural network classifiers based on pretrained transformer models, and Carmen[4] to infer the geolocation—at the United States state level—of the tweets that the classifier predicted as potential cases.
- End-to-end automatic NLP pipeline:



Results

- The classifier based on the BERT-Base-Uncased[5] pretrained model achieved an F_1-score of 0.70 (precision = 0.72, recall = 0.67), and the classifier based on the COVID-Twitter-BERT[6] pretrained model achieved an F_1-score of 0.76 (precision = 0.76, recall = 0.76).
- We deployed our automatic pipeline, using the COVID-Twitter-BERT classifier, on more than 85 million unlabeled tweets continuously collected from the Twitter Streaming API between March 1 and August 21, 2020, detecting 13,714 "potential case" tweets for Carmen could infer a United States state-level geolocation.
- We automatically detected "potential case" tweets from all 50 states, with the highest numbers posted in California, New York, Texas, and Florida. These tweets include reports of potential cases 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.



Acknowledgments

- This work was supported by the National Institutes of Health (NIH) National Library of Medicine (NLM) grant R01LM011176 and National Institute of Allergy and Infectious Diseases (NIAID) grant R01AI117011. We would like to thank Alexis Upshur for contributing to tweet annotations.

1 Menni C, Valdes AM, Freidin MB, Sudre CH, Nguyen LH, Drew DA, Ganesh S, Varsavsky T, Cardoso MJ, El Sayed Moustafa JS, Visconti A, Hysi P, Bowyer RCE, Mangino M, Falchi M, Wolf J, Ourselin S, Chain AT, Steves CJ, Spector TD. Real-time tracking of self-reported symptoms to predict potential COVID-19. Nat Med. 2020;26(7):1037-1040.
2 Lauer SA, Grantz KH, Bi Q, Jones FK, Zheng Q, Meredith HR, Azman AS, Reich, NG, Lessler J. The incubation period of Coronavirus Disease 2019 (COVID-19) from publicly reported confirmed cases: estimation and application. Ann Intern Med. 2020;172(9):577-582.
3 Klein AZ, Cai H, Weissebacher D, Levine LD, Gonzalez-Hernandez G. A natural language processing pipeline to advance the use Twitter data for digital epidemiology of adverse pregnancy outcomes. Journal of Biomedical Informatics: X 2020;8:100076.
4 Drezde M, Paul M, Bergsma S, Tran H. Carmen: a Twitter geo-location system with applications to public health. In: Proceedings of AAAI 2013 Workshop Expanding the Boundaries of Health Informatics Using Artificial Intelligence; 2013. p. 20-24.
5 Devlin J, Cheng MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of NAACL-HLT; 2019. p. 4171-4186.
6 Müller M, Salathé M, Kummervold P. COVID-Twitter-BERT: a natural language processing model to analyse COVID-19 content on Twitter. arXiv 2005.07503 [Preprint]. 2020. Available from: https://arxiv.org/abs/2005.07503.

Scan to download the full paper and access the data sets: