Automatically Identifying Twitter Users for PrEP-Related Interventions

Ari Z. Klein[1], Steven Meanley[2], Karen O'Connor[1], José A. Bauermeister[2], Graciela Gonzalez-Hernandez[1]
[1] Health Language Processing Center (https://healthlanguageprocessing.org),
Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania
[2] Department of Family and Community Health, University of Pennsylvania
{ariklein, gragon}@pennmedicine.upenn.edu

Background

- Pre-exposure prophylaxis (PrEP) is highly effective at preventing the acquisition of HIV[1]. There is a substantial gap, however, between the number of people in the United States who have indications for PrEP and those who are prescribed PrEP[2].
- Although Twitter content has been analyzed as a source of PrEP-related data (e.g., barriers), methods have not been developed to enable the use of Twitter as a platform for implementing PrEP-related interventions.

Objectives

- Men who have sex with men (MSM) are the population most affected by HIV in the U.S.[3] Therefore, the objectives of this study were to (1) develop an automated natural language processing (NLP) pipeline for identifying men in the U.S. who have reported on Twitter that they are gay, bisexual, or MSM; and (2) assess the extent to which they demographically represent MSM in the U.S. with new HIV diagnoses.

Conclusions

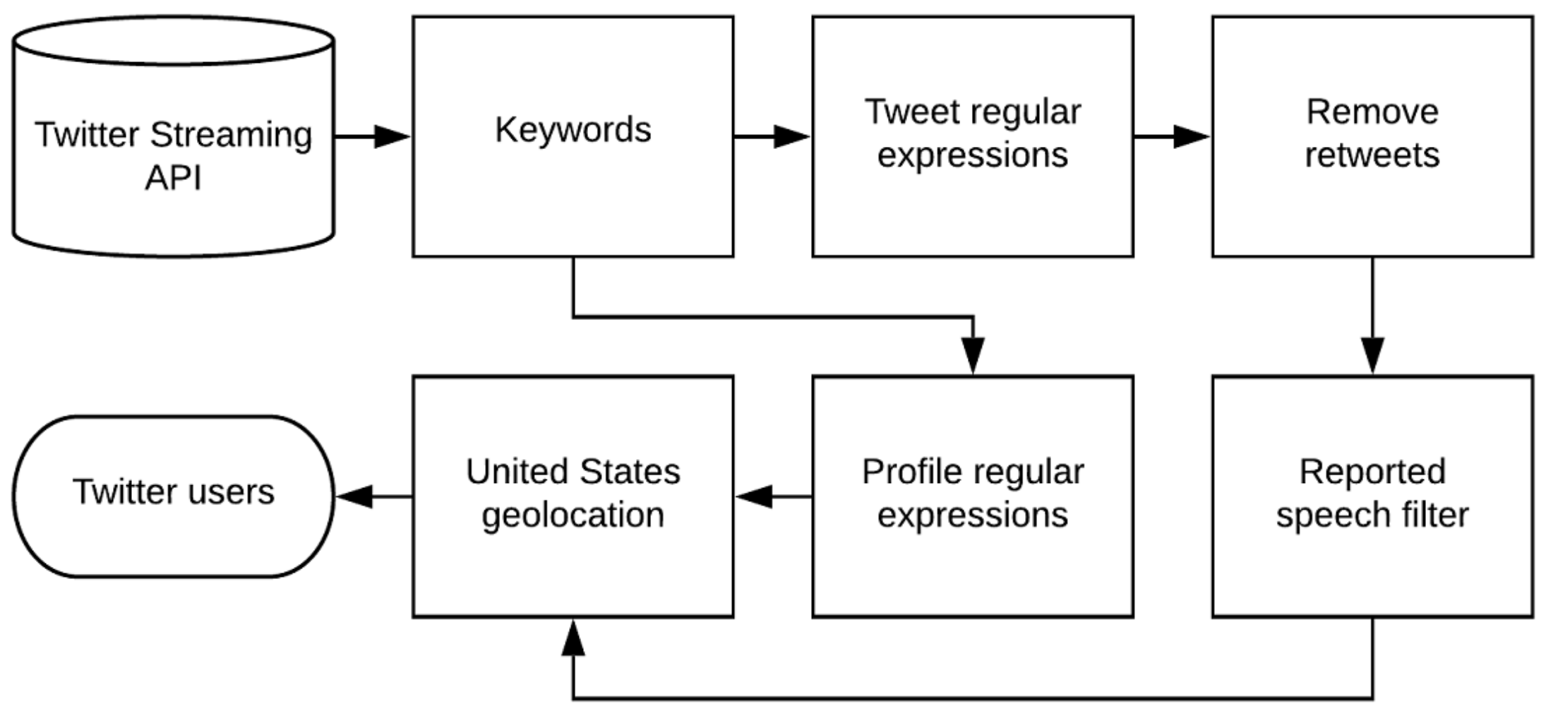
- Our automated pipeline can be used to identify MSM who are largely in the regions and age groups most affected by HIV in the U.S., laying the groundwork for using Twitter on a large scale to directly target PrEP-related interventions at this population.

Acknowledgments

- This research was supported by a grant from the Penn Center for AIDS Research (CFAR), an NIH-funded program (P30 AI 045008). We thank Ivan Flores for contributing to software development and Alexis Upshur for contributing to tweet annotations.

Methods

- *Pipeline*. Between September 2020 and January 2021, we identified 10,043 users.



Twitter Streaming API → Keywords → Tweet regular expressions → Remove retweets → Reported speech filter → Profile regular expressions → United States geolocation → Twitter users

- *Evaluation.* Two annotators distinguished true- and false-positive self-reports in a sample of 500 tweets and 500 profiles of the 10,043 identified users.

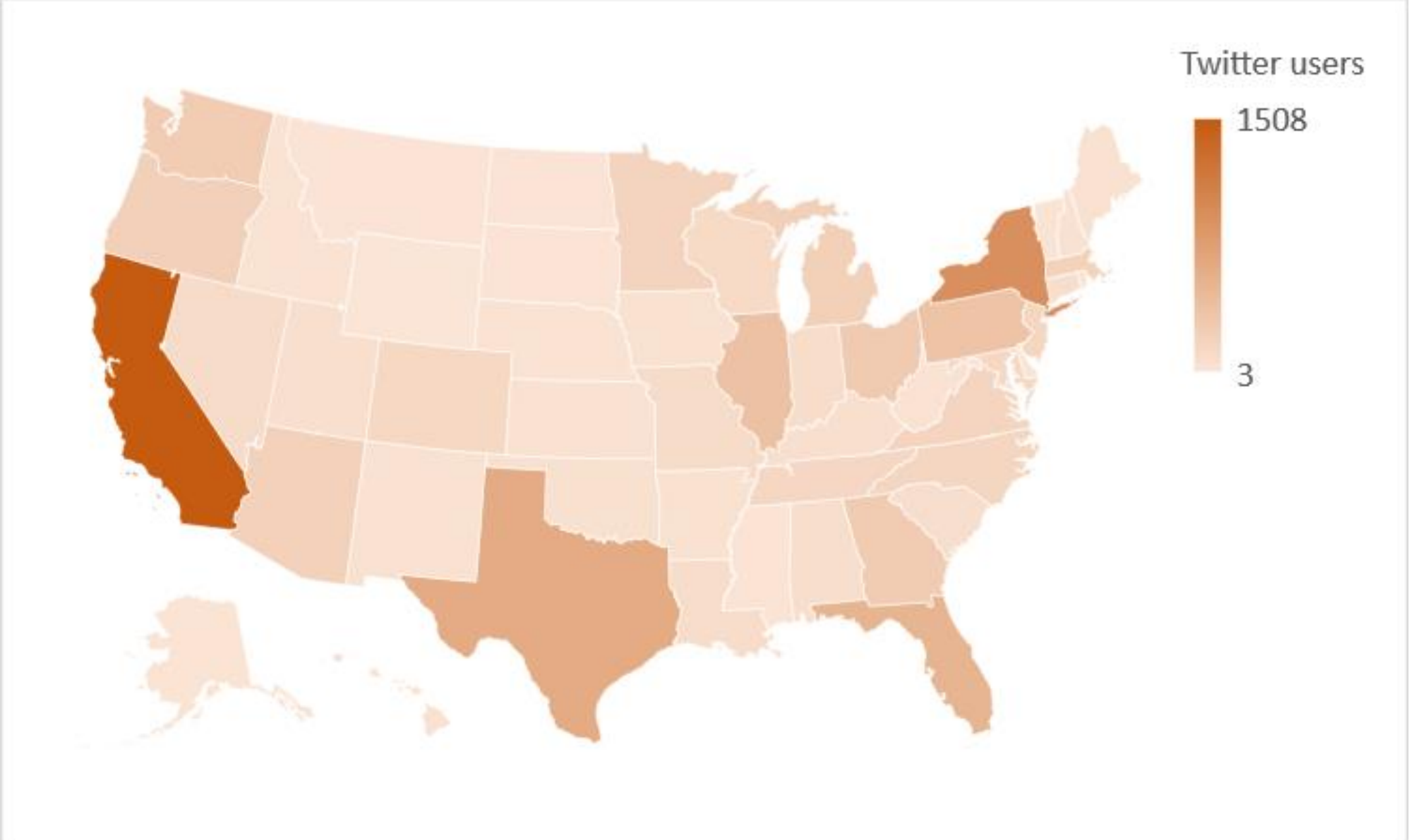
Tweet	End the FDA's discriminatory and unscientific policy against gay men like me donating blood.	+
Tweet	Today, we remember Matthew Shepard who's life was cut short as a result of a hate crime due to his identity as a gay male.	-
Profile	A proud black gay guy.	+
Profile	50+ gay trans man, writer, film and food lover. He/him OR they/them.	-

- *Demographics.* We used Carmen[4] and ReportAGE[5] to automatically identify the geolocation and age, respectively, of the 10,043 identified users.

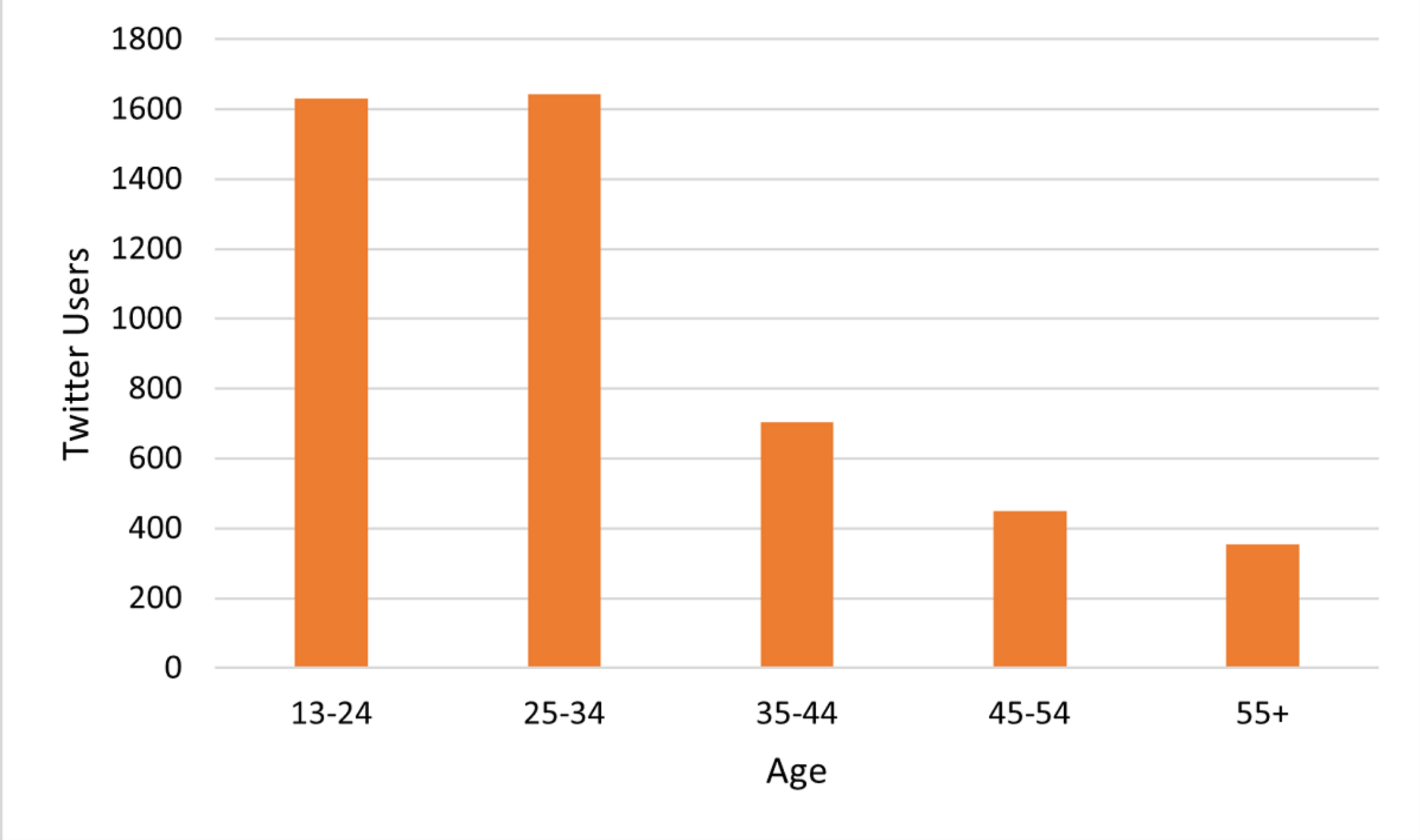
Scan to download a preprint of the full paper, which is in press at *JMIR Public Health and Surveillance*:

Results

- *Evaluation.* Among 1,000 of the 10,043 identified users, 417 (83.4%) of their 500 tweets and 430 (86%) of their 500 profiles were annotated as true positives, establishing an overall precision of 0.85.
- *State-level geolocation.* We detected a state-level geolocation for 8,756 (87.6%) of the 10,043 identified users. Among these 8,756 users, 5,096 (58.2%) were in the 10 states with the highest numbers of new HIV diagnoses[3].



- *County-level geolocation.* We detected a county-level geolocation for 6,240 (71.2%) of the 8,756 users for which we detected a state-level geolocation. Among these 6,240 users, 4,252 (68.1%) were in counties or states considered priority jurisdictions by *Ending the HIV Epidemic*[6].
- *Age.* We detected an age of ≥13 years for 4,782 (47.6%) of the 10,043 identified users. The majority of the users are in the same two age groups as the majority of MSM with new HIV diagnoses.



[1] Grant RM, Lama JR, Anderson PL, McMahan V, Liu AY, Vargas L, Goicochea P, Casapía M, Guanira-Carranza JV, Ramirez-Cardich ME, Montoya-Herrera O, Fernández T, Veloso VG, Buchbinder SP, Chariyalertsak S, Schechter M, Bekker LG, Mayer KH, Kallás EG, Amico KR, Mulligan K, Bushman LR, Hance RJ, Ganoza C, Defechereux P, Postle B, Wang F, McConnell JJ, Zheng JH, Lee J, Rooney JF, Jaffe HS, Martinez AI, Burns DN, Glidden DV, iPrEx Study Team. Preexposure chemoprophylaxis for HIV prevention in men who have sex with men. N Engl J Med. 2010;363(27):2587-2599.
[2] Smith DK, Van Handel M, Wolitski RJ, Stryker JO, Hall HI, Prejean J, Koenig LJ, Valleroy LA. Vital signs: estimated percentages and numbers of adults with indications for preexposure prophylaxis to prevent HIV acquisition—United States, 2015. MMWR Morb Wkly Rep. 2015;64(46):1291-1295.
[3] Centers for Disease Control and Prevention. Diagnoses of HIV infection in the United States and dependent areas, 2018 (updated). URL: https://www.cdc.gov/hiv/library/reports/hiv-surveillance/vol-31/index.html
[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] Klein AZ, Magge A, Gonzalez-Hernandez G. ReportAGE: automatically extracting the exact age of Twitter users based on self-reports in tweets. PLoS One. 2022;17(1):e0262087.
[6] Centers for Disease Control and Prevention. Ending the HIV epidemic. URL: https://www.cdc.gov/endhiv/jurisdictions.html