Automatic Cohort Determination from Twitter for HIV Prevention amongst Black and Hispanic Men

Davy Weissenbacher, PhD[1], J. Ivan Flores, B.S.[1], Yunwen Wang, M.A., M.S.[2], Karen O'Connor, M.S.[1], Siddharth Rawal, B.S.[1], Robin Stevens, PhD, MPH[2], Graciela Gonzalez-Hernandez, PhD[1]
[1]University of Pennsylvania, Philadelphia, Pennsylvania, USA;
[2]University of Southern California, Los Angeles, California, USA

Abstract

Recruiting people from diverse backgrounds to participate in health research requires intentional and culture-driven strategic efforts. In this study, we utilize publicly available Twitter posts to identify targeted populations to recruit for our HIV prevention study. Natural language processing and machine learning classification methods were used to find self-declarations of ethnicity, gender, age group, and sexually-explicit language. Using the official Twitter API we collected 47.4 million tweets posted over 8 months from two areas geo-centered around Los Angeles. Using available tools (Demographer and M3), we identified the age and race of 5,392 users as likely young Black or Hispanic men living in Los Angeles. We then collected and analyzed their timelines to automatically find sex-related tweets, yielding 2,166 users. Despite a limited precision, our results suggest that it is possible to automatically identify users based on their demographic attributes and Twitter language characteristics for enrollment into epidemiological studies.

1 Introduction

Recruiting people from diverse backgrounds to participate in health research requires intentional and culture-driven strategic efforts. New media and technology provide novel opportunities to reach diverse populations. In recent decades, online networking sites have gained increased popularity as research recruitment tools, e.g. using Facebook to recruit young adults for inspecting their alcohol and drug use[1] or young smokers for a cessation trial[2] (see a review[3]). While social media can in theory increase the reach of recruitment efforts and health promotion messaging, it does not ensure the equity of reach. Social media recruitment and interventions can reproduce long-standing health disparities if those efforts fail to target diverse groups based on their varying media use patterns and digital access.

Geolocation data has been used for public health surveillance and has been used in conjunction with temporal, textual, and network data to monitor e-cigarette use[4], opioid abuse[5,6], influenza[7], and HIV-related social media discussions[8–10]. However, many of these applications remain largely descriptive of the social media content itself, rather than predictive or proactively applied to public health practice.

In consort with these geolocation approaches, we can utilize additional social media user data to identify target groups that meet specific demographic and behavioral characteristics for recruitment. A targeted social-media-based recruitment approach can identify communities in greatest need and can overcome the challenges of traditional recruitment approaches. Integrating the geographical location of social media users with content-based attributes, such as what we propose, provides a valuable tool for initial triage that can unobtrusively and inexpensively help identify potential study participants based on these detectable attributes, within the validated performance limitations of the automatic methods.

In this study, we utilize publicly available social media posts to identify predetermined populations to recruit for our HIV prevention study, targeting specific ethnic, gender, and age groups. The open user-generated content on Twitter was used to identify distinct subgroups of users who possibly meet the study's selection criteria and can be purposefully contacted. Compared to the undifferentiated social media postings that are likely confined by personal networks, our approach aims to better unleash the potential of large-scale social media data, to more efficiently find a larger number of the target population that is otherwise missed through traditional approaches. Specifically, this study aims to identify the likely members of ethnic (Black/Hispanic), gender (male), and age (18-24) groups in a particular geographic location (Los Angeles, California), whom we will recruit as the participants for an HIV prevention intervention. In this paper, we describe how we used geolocation techniques to bound the sampling frame to a digital sphere of people

living in a specific urban space. We then used novel Natural Language Processing (NLP) approaches to identify users' demographic characteristics and behaviors based on the content of their profiles and tweets.

The main contributions of our work are the collection of two corpora, (1) a corpus of 2,577 Twitter users profiles annotated with the demographic information available in their profiles, (2) a corpus of 3,500 tweets annotated with the use of sexually-explicit language and likely posted by young Black and Hispanic men living in Los Angeles, and (3) the release of a classifier trained to identify such language. Our annotation guidelines and the code of our classifier are available at *https://bit.ly/33nF2p3*.

2 Related Work

Health professionals are increasingly integrating social media in their daily activities, not only to disseminate new knowledge and extend their professional network but also to communicate with patients and collect health-related data[11]. The possibility to identify and recruit participants for research studies using social media is one of its most appealing promises for health research. As trials continue to be canceled or discontinued due to poor quality recruitment through traditional methods[12], recruitment through social media appears to be a good complement, or a better alternative to reach underserved populations.[13,14].

Although targeted advertising on Facebook has been used for recruiting participants[14], we opted for Twitter because it offers several advantages for our study. Twitter has a large base of active users[15], a free application programming interface (API) allowing for simple, real-time access to the data, and, due to its nature, it represents a default media for the users to publicly share and for researchers to have open access to their expressed ideas, activities, and beliefs; enabling selection via what they post and from where they post rather than only passively waiting for them to click on an ad. These makes Twitter a valuable recruitment tool that can complement other social media platforms[16]. For example, Twitter helped to recruit cancer survivors in[17], young cigarillo smokers in[18], or users of direct-to-consumer genetic testing services[19].

Still, the most common method to recruit participants with Twitter is to use it as a broadcast platform. For example, researchers create a web page to describe their study and through their personal Twitter accounts, or an account dedicated to the study, advertise the study by frequently posting links to the web page. They may also ask colleagues, relevant organizations, or influential actors in their network to tweet/retweet the links to the study to reach more users. This approach relies on existing social networks and can lead to the exclusion of users in demographic groups outside of those research networks. As done in traditional recruitment through TV or local journals, researchers may also buy ads to promote their study among potentially eligible users as identified by the marketing tools of the platform. While this may be effective for some studies, this approach is passive: researchers advertise their studies and wait for interested individuals to contact them.

Few researchers have used the data posted by the users themselves to identify eligible users and directly contact them to join their study[20]. In a seminal work, Krueger et al.[21] identified transgender people communicating on Twitter by collecting tweets mentioning 13 transgender-related hashtags in order to examine their health and social needs. Hashtags allow for the collection of tweets of interest but they have strong limitations. Amongst these, we note: (a) 58% of the tweets collected in their study were irrelevant to the transgender community and had to be manually excluded, (b) the hashtags used by a community change quickly over time, and (c) all users of interest not using the hashtags were missed. The protocol published by Reuter et al.[22] proposed a similar approach, but complemented the hashtags with keywords to detect patients with cancer on Twitter. The results of their study are, at the time of writing, under peer-review[23] but they indicate the feasibility of this approach using a combination of manual curation and automated processes. In this study, we will follow an approach that builds on our prior work[24] where we successfully combined keywords and machine learning to detect women publicly announcing their pregnancy.

3 Methods

We aimed to find on Twitter potential users to consent for inclusion in our study: young men (aged 18 to 24) who are Black and/or Hispanic and live in the metropolitan area of Los Angeles, California, USA and who publicly post sex-related content. We summarize our process in Figure 1. The following sections detail each step.

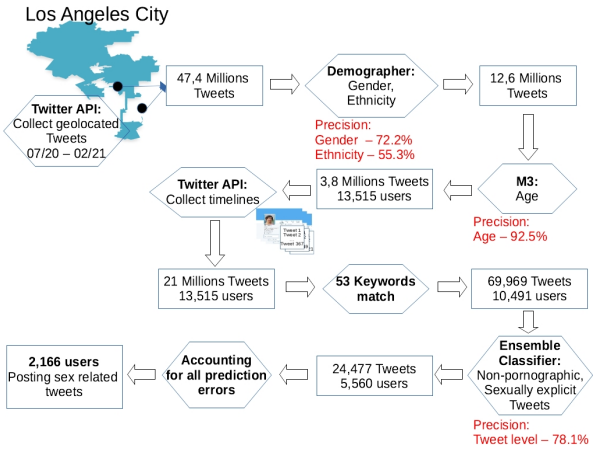

Figure 1: Cohort detection process on Twitter.

3.1 Geographical Locations and Demography Detection

To detect our target population, we first collect tweets posted in the Los Angeles (LA) area, using the official Twitter Search RESTful API. We defined two circular areas covering the city of LA. The center of the first circle was positioned at latitude 33.987816 and longitude -118.313296, and the center of the second circle at latitude 33.877130 and longitude -118.21004. Both circles had a radius of 16km. These circles roughly cover the areas of Compton and South-Central LA which are neighborhoods with predominantly Black and Hispanic populations[25]. The free API facilitates collecting tweets matching the given geographic criteria posted up to seven days before the query. We collected samples of 200,000 tweets posted in these areas every 6 days between July 06, 2020, until February 26, 2021. After removing duplicate tweets, our initial collection included 47.4 million tweets.

We then applied two filters sequentially, based on results from the existing classifiers Demographer[26] and M3[27], on the 47.4 million tweets initially collected, shown in Figure 1. Among competing tools[28], we chose Demographer and M3 based on their availability and performance. The first filter, which consists of rules on the output by Demographer, helped us remove tweets not posted by Black or Hispanic male users. Given a single tweet, Demographer searches for the given name of the user in the profile or the username. It distinguishes individuals from organizations by relying on neural networks to learn morphological features from the name and higher-level features from the user's profile. When a user is identified as an individual, the gender and race or ethnicity of the person are inferred by the software. We removed 34.9 million tweets (73.5%) posted by users that were flagged as either institutional accounts or not meeting the gender and race/ethnicity inclusion criteria (Black or Hispanic men) from our initial collection, leaving a total of 12.6 million tweets to be processed by the second filter.

The second filter uses rules on the output of M3 to remove tweets posted by users younger or older than the target age bracket of 19-29 years old. Although our original study calls for users aged 18-24, 19-29 is the closest age bracket computed by the tool. M3 relies on a deep neural network to jointly estimate the age, gender, and individual-/organization-status of the user from their profile images, usernames, screen names, and the short descriptions in their profiles. We removed 8.8 million tweets (70.2%) posted by users outside of the target age bracket, leaving a total of 3.8 million tweets from 13,515 distinct users of interest: men, aged 19 to 29, that are likely Hispanic and/or Black and live in the LA area.

In order to validate the two initial filters (Demographer for gender and race/ethnicity prediction and M3 for age prediction), we randomly selected and annotated 2,577 users' profiles (19%) from the 13,515 users. In agreement with best practices for deriving race/ethnicity information from Twitter datasets[29], three annotators (one graduate student and two undergraduates) that match the target demographic (young Black and/or Hispanic) were trained using the guidelines developed for the task[1]. We provided annotators the profile information of users directly obtained through the Twitter API including the username, screen name, profile description, and listed location. We also allowed the

[1] Available at https://bit.ly/33nF2p3

annotators to search the user's Twitter profile directly, including other social media platforms or websites that the user provided links to in their Twitter profile, to help establish the information needed to validate the automatic prediction of user demographics.

For each of the 13,515 users identified by our pipeline, we collected up to 3,000 of their latest tweets, which is the maximum allowed through the Twitter API. This yielded 21 million tweets (4.35G) to be analyzed for our last aspect of the selection criteria: identify users who post tweets with sex-related language. The next section details this last classifier, which we developed specifically for this study.

3.2 Sex-related Language Detection

We developed an ensemble of classifiers to identify users posting tweets with sex-related language in our collection. Sex-related content included tweets about sexual intercourse, sexual desire, and sexual behavior. We excluded tweets focused solely on genitalia. We trained our ensemble with supervision, the most effective training approach, to learn a binary function: for a given tweet, our ensemble returns 1 if the content of the tweet is sex-related, 0 otherwise.

We used an existing corpus of 11,175 annotated tweets to pre-train our classifiers, as pre-training has often been found to improve the training phase. The corpus is described in a recent study[10]. The authors randomly collected tweets from the 1% of publicly available tweets posted between January 1, 2016, and December 31, 2016. They isolated 6,949 tweets posted by young men living in the US by using existing tools and a list of keywords related to sex and HIV, such as *HIV* or *condoms*, among many. They manually annotated a subset of 3,376 tweets as pornographic (759, 22.5%), sex-related (1,780, 52.7%), and not sex-related tweets (837, 24.8%). In their study, they classified pornographic tweets based on a tweet format that included extensive use of hashtags and links to other websites. These hashtags were often of several sexually explicit words. They estimated the inter-annotator agreement to have an intraclass correlation of 91.2%. They later extended their corpus by selecting 7,799 additional tweets by following the same process and annotated the tweets as either "not pornographic but sex-related" or "not sex-related nor pornographic". For our study, we used all of the annotated tweets, that is, 11,175 tweets (3,376 + 7,799 tweets).

We then fine-tuned and evaluated our classifier on a second, new, corpus sampled from the 21 million tweets posted by our population of interest, the 13,515 users identified by our pipeline. We selected a subset of candidates likely to be sex-related tweets by searching in the 21 million tweets mentions of the 53 keywords or emojis that have a sexual connotation from the list used in the previous study[10]. This search returned 69,969 tweets posted by 10,491 users. We randomly sampled 3,500 (about 5%) of the tweets for annotation. Of these, 1,164 tweets were tagged as having sex-related language (but not pornographic) and 2,336 were tagged as not sex-related or pornographic. The inter-annotator agreement was substantial with 0.652 Cohen Kappa score. We found very few pornographic tweets among the 69,969 tweets because of the selection of our filters which identify individual users living in LA and filtered out organization and bot accounts, likely to be the main sources of pornographic tweets.

Our ensemble is composed of recurrent neural networks and transformers classifiers since these models are state-of-the-art and have been shown to encode efficiently the semantic content of sentences. We chose well-known architectures and available embeddings: 1) a Bidirectional Long Short-Term Memory (biLSTM) network which inputs are static word embeddings pre-trained on 2 billion tweets with GloVe; 2) a biLSTM which inputs are contextualized BERT embeddings (Bidirectional Encoder Representations from Transformers); 3) a single feedforward network which inputs are BERT embeddings; 4) another single feedforward network which inputs are RoBERTa embeddings (Robustly optimized BERT approach)[30]; 5) a biLSTM which inputs are RoBERTa embeddings.

Given the relatively small size of our corpora, we evaluated our ensemble of classifiers with 5-fold cross-validation, per accepted evaluation standards. For each fold, we pre-trained each classifier on the existing corpus of 11,175 tweets, we randomly split our corpus of 3,500 tweets into three sets, keeping 70% of the data for training, 10% for validation, and 20% for evaluation. During training, we fine-tuned the pre-trained transformer models with an AdamW optimizer, setting the learning rate to 4e-5 and a batch size of 8. For the biLSTM classifiers, we used default parameters of the Flair NLP library: a Stochastic Gradient Descent (SGD) optimizer with a learning rate of 0.1 and a batch size of 32. We trained the models during 5 epochs and evaluated them at each epoch on the validation set. We selected the models at the epochs where they achieved the highest accuracy on the validation set for testing. We used weighted

average voting to ensemble the decisions of the classifiers on the test set, with the weight of each classifier being their respective performance on the validation set. We evaluated our classifiers using the standard metrics of precision and recall for classification, as well as their harmonic mean, the F1-score.

3.3 Accounting for Prediction Errors in our Cohort Collection

Although we applied state-of-the-art methods to automatically predict the demographic features of Twitter users and the linguistic content of their posts, our predictions remain noisy. We estimated the real number of users detected by our pipeline by correcting our numbers using the precision of each component in our pipeline, calculated using our manual annotation of the 3,500 tweets sampled from our corpus. With these performance metrics (precision), we estimated how many users eligible for our study were actually amongst the 5,560 resulting after running our complete pipeline (see the last steps in Figure 1 and the Discussion section for details).

4 Results

4.1 Geographical Locations and Demography

We first assessed agreement on gender validation. Prior to calculating agreement, any user whose account was removed at the time of annotation for any one of the annotators was removed from the corpus. In total, 2,577 unique user profiles were validated by at least one annotator. All agreement measures were calculated using the IRR package in R (version 4.0.0). Agreement was calculated using Cohen's kappa for two raters and Fleiss kappa for three raters. Table 1 shows the results of the agreement calculations.

Annotator group	A+B	B+C	A+C	A+B+C
Number of users	720	382	382	382
Percent agreement	77.4	78.3	90.3	74.1
Cohen's κ	0.568	0.575	0.776	—
Fleiss κ	—	—	—	0.618

Table 1: Inter-annotator agreement for gender determination.

Disagreements were resolved by taking the majority class for users that were validated by three annotators. In the event that there was no majority, and for disagreements on users validated by only two annotators, the annotation was adjudicated by the developer of the guidelines. Table 2 shows the results of validation after adjudication.

Annotation set	Number of users	Account removed (%)	Female (%)	Not a personal account (%)	Unable to determine	Male (%)
Triple	400	20 (5.0)	66 (16.5)	14 (3.5)	22 (5.5)	278 (69.5)
Double	347	22 (6.3)	73 (21.0)	11 (3.2)	26 (7.5)	215 (62.0)
Single	1830	44 (2.4)	349 (19.1)	14 (0.7)	117 (6.4)	1306 (71.4)
Total	2577	86 (3.3)	488 (18.9)	39 (1.5)	165 (6.4)	1799 (69.8)

Table 2: Results of gender validation.

Agreement for race/ethnicity annotations was then calculated for the 1,799 users validated as male. Before calculating agreement, all annotations were normalized to three groups, 'y' if the annotator identified the user as Black and/or Hispanic, 'n' for any other race/ethnicity identified, and 'u' for those the annotator was unsure. Table 3 shows the agreement measures for race/ethnicity.

Annotator group	A+B	B+C	A+C	A+B+C
Number of users*	441	241	274	240
Percent agreement	76.9	71.8	71.5	58.8
Cohen's κ	0.543	0.459	0.545	—
Fleiss κ	—	—	—	0.449

Table 3: Inter-annotator agreement for race/ethnicity determination.

*Note. * The number of users differs for each group due to missing annotations, based on the guidelines if an annotator determined a user to be non-male, they did not annotate other demographic information.*

Adjudication for race/ethnicity was performed in the same way as was done for gender. Table 4 shows the results after

adjudication. Note that the accounts removed were those not found when adjudicating the disagreements and those accounts were removed from the corpus.

Annotation	Number of male users	Black and/or Hispanic (%)	Not Black and/or Hispanic (%)	Unable to determine (%)	Account removed
Triple	278	166 (59.7)	55 (19.8)	56 (20.1)	1
Double	215	133 (61.7)	61 (28.4)	18 (8.4)	3
Single	1306	694 (53.1)	391 (29.9)	221 (16.9)	0
Total	1799	993 (55.2)	507 (28.2)	295 (16.4)	4

Table 4: Results of race/ethnicity validation.

The annotators derived location exclusively from the location field provided in the profile, therefore the agreement was very high (95.8%) for location annotations. Of the 993 users, 982 (98.9%) had information in the location field that the annotators used for validation. After adjudication and normalization of the names of neighborhoods and cities in Los Angeles County, there were 26 locations annotated as outside LA county, 8 were due to the user entering a non-place name in the location field, 17 were locations outside the target area, and 1 was a missed annotation.

Our annotators only used the descriptions in the user profiles or information provided on other linked sites to validate their age; however, age information was sparse. Of the 993 users validated as male and Black or Hispanic, only 67 (6.7%) had an age annotated by one of the annotators; 62 (92.5%) of the 67 users were in the 19-29 year-old age range. We are currently validating the age by crossing multiple dimensions of the users' profiles: searching for self-declaration of ages in the timelines, looking at the photos posted by the users, and searching for the ages of other members of the networks the users frequently interact with.

4.2 Detecting sex-related tweets

We report the performance of the ensemble as well as the performance of each classifier composing the ensemble in Table 5. With an average of 0.7814 F1-score during the 5-fold cross-validation ($SD = 0.019$), the ensemble achieved the best performance. We checked if the disagreements between the predictions of the ensemble and those of the best performing individual classifier, Roberta + biLSTM, were statistically significant. For our computation, we used the predictions of the fold where their F1-scores were the closest. With a McNemar test, we found the proportions of disagreements between the two classifiers too close to rejecting the null hypothesis (with α=0.1). We also evaluated the performance of our ensemble when trained and evaluated only on the 3500 tweets of our corpus, that is, when none of the classifiers composing the ensemble were pre-trained on the existing corpus of 11,175 annotated tweets. With an average of 0.7596 F1-score during the 5-fold cross-validation, the ensemble without pre-training performed slightly worse than the ensemble trained on all available tweets. We compared the predictions of both ensembles during the fold where their F1-scores were the closest with a McNemar test. Whereas the ensemble trained with all examples achieved better performance than the other ensemble during all folds, we found the proportions of disagreements between the two ensembles too close to rejecting the null hypothesis (with α=0.05), showing a marginal improvement of the pre-training for our task.

Classifier	**P**	**R**	**F1**
BERT + FF	0.7138	0.6984	0.7044
RoBERTa + FF	0.7221	0.8003	0.7586
GloVe + biLSTM	0.6661	0.6940	0.6772
BERT + biLSTM	0.7111	**0.8170**	0.7601
RoBERTa + biLSTM	0.7283	0.8012	0.7619
Ensemble (pre-trained)	0.7810	0.7825	**0.7814**
Ensemble (no pre-training)	**0.7908**	0.7322	0.7596

Table 5: Performance of classification of non-pornographic sex-related tweets.

We analyzed the false positive predictions of our ensemble during the k-fold where it achieved its best precision score of 0.8206, and 0.7821 Recall, 0.8009 F1-score on our test set. The classifier made 40 false positives (FP) predictions. Most FPs, 25% (10/40), were tweets not talking about sex but the keywords or phrases occurring in the tweets could have sexual meanings in other contexts. Closely followed tweets with insults or swearing 22.5% (9/40) or jokes

15% (6/40) using sexual references. These tweets have explicit phrases but their main messages were not sexually related. Also, 12.5% (5/40) were tweets with sexual concepts used as elements of comparison in metaphors; 7.5% (3/40) were tweets posted by users speaking about fictional characters, masturbation, or describing their genitals, which our guidelines excluded. We found that only 7.5% (3/40) tweets were related to pornography and incorrectly labeled by our ensemble, including two tweets posted by sex workers self-promoting their content, and a third tweet as an advertisement written in a descriptive style very close to the style used by users commenting on pornographic materials. In the end, 7.5% (3/40) were deemed annotation errors by the final adjudicator. The decisions computed by neural networks are difficult to explain, as a result, we were unable to identify the reason for the misclassification of the last remaining tweet by our model.

We also analyzed the 51 false negatives (FN) predictions of the classifier. Most FNs, 35.2% (18/51) were tweets where sexual statements occurred in long sentences discussing a topic not sexually related; 23.5% (12/51) tweets were sexual meanings suggested often as jokes; 9.8% (5/51) tweets were tweets where a sexual keyword occurred but the tweets were very short - less than five words - providing limited context for the ensemble to determine their meaning. Finally, 7.8% (4/51) were tweets written with phonetic spellings or not written in English. We were unable to identify the reasons for the misclassification of the remaining 23.5% (12/51) tweets.

4.3 Detecting users posting sex-related tweets

Taking all 21 million tweets by the 13,515 users identified by our pipeline as Black or Hispanic men that are 19-29 years old and live in the Los Angeles area, and matching our list of keywords related to sex in their timeline and classifying the tweets detected with our ensemble, there were 5,560 users at the end of our entire pipeline. We observed that users are likely to post several sex-related tweets, with 58% (3211/5560) posting more than 3 such tweets. Whereas our ensemble did not detect perfectly single sex-related tweets, the fact that users post sex-related tweets more than once allows us greater precision when estimating the likelihood of a user having posted such Tweets. We grouped the 5,560 users according to the total number of sex-related tweets posted as predicted by the ensemble. We modeled the detection of users posting sex-related tweets as a binomial experiment, where a success is the ensemble correctly identifying a sex-related tweet. Since the precision of our ensemble was estimated to be 0.781, out of the 453 users having posted only 1 tweet predicted as sex-related we can assume that the ensemble correctly detected 353 users (453-(453*0.219)). Out of the 503 users having posted exactly 2 tweets predicted as sex-related , we estimated that the ensemble correctly detected 478 users (503-(503*0.048)) since the probability of our ensemble to be wrong on both tweets predicted as sex-related is 0.048 (with B(n=2, p=0.781) and k=0, where n is the number of predictions of the ensemble, p its precision and k the number of time the ensemble was correct). Following similar reasoning, we estimate that we have correctly detected 448 out of 453 users who posted exactly 3 tweets. We assume that the remaining 4,151 users who posted 4 or more tweets predicted as sex-related by our ensemble were correctly detected since the probability of k=0 was lower than 0.0025. In total, we estimate that among the 5,560 users, our ensemble was able to correctly identify 5,430 users posting sex-related tweets, indicating a high score of 0.977 precision.

Discussion

From a sample of 2,577 users' timelines that we manually inspected for validation, we were able to identify 993 Black or Hispanic men most likely to live in Los Angeles; the validation of the age of the users is still ongoing. With a limited score of 39.9% precision ($993/(2577-90)$), automatic tools for inferring Twitter users' demographic information are still in their infancy and require improvement. Yet, despite the imprecision and sparsity in the demographic attributes of Twitter users, we were able, using these tools, to automatically identify around 5,392 (13,515*0.399) users who are really from the population of interest. Given we will reach out to the 13,515 users for consent and enrollment, any errors in the automatically derived demographic information will be evident and the users excluded if needed when verifying their eligibility for our HIV prevention study. In future work, we will improve the detection of errors when determining the gender and ethnicity attributes. In this study, we only relied on the Demographer to compute these attributes, whereas both Demographer and M3 distinguish individuals from organizations and, for individuals, deduce their gender. We will integrate both tools in our pipeline and submit for manual verification the users where the tools disagree. We will also search for self-declarations of these attributes in the timelines with regular expressions, a dimension that has not been exploited by the previous tools.

Our ensemble achieved a moderate performance when detecting sex-related tweets with 0.781 F1-score, however, since users are most likely posting multiple sex-related tweets, we found its precision to be high when detecting users posting multiple sex-related tweets with a score of 0.977. Among the 13,515 users of interest our ensemble selected 5,430 users posting sex-related tweets. Accounting for the errors in the predictions of the demographic of the users, we assume that among these 5,430 users, 2,166 (5430*0.399) are actually Black or Hispanic men in the Los Angeles Area, given only 39.9% of the users were identified as such in the set of 2,577 user profiles manually inspected.

In this study, we designed our pipeline to be precise because we have limited resources to contact users and validate their demography. A few changes can be made in our approach to increase the size of our cohort (semi-)automatically. First, we could replace the Twitter's streaming API with the Twitter's firehose to collect tweets posted during a given period. Second, we could extend semi-automatically the list of sexually connoted keywords by looking for and including in our list all sex-related phrases frequently occurring in the tweets predicted by our ensemble. Last, we did not store a large number of tweets from our initial collection of 47.4 million tweets based on the assumption that the users were not Black or Hispanic users. However, we found in our results that the ethnicity was imprecisely inferred automatically with a low 55.3% precision score. In future research, we may remove the ethnicity filter from the pipeline and solely rely on the geolocations of the tweets, given they originate within two neighborhoods with predominantly Black and Hispanic populations, which may largely increase the number of users detected.

We acknowledge the limited applicability of the approach to populations with non-binary gender identities. The concept of "gender" encompasses biological, psychological, social, and cultural factors[31]. Our approach, as primarily estimating what appears to be the Twitter user's biological sex (female/male), presents limitations in capturing the nuanced social construction and self-identification of gender. During manual verification of gender prediction, we encountered users that are non-binary or present themselves as women despite the classifier's prediction of being men. In future research, we will explore ways to better include sexual and gender minority populations.

5 Conclusion

In this study, we aimed to automatically identify young Black or Hispanic men living in Los Angeles who use sex-related language publicly on Twitter. It was important to narrow the target cohort as much as possible as we intend to recruit them into the HIV prevention study that is the focus of our grant (NIDAR21DA049572). We used the official Twitter API and available tools, Demographer[26] and M3[27], to extract the age, gender, and ethnicity of users tweeting from Los Angeles. In addition, we developed an ensemble of neural networks to detect sex-related language. Our ensemble achieved a 0.7814 F1-score on our test set when identifying individual tweets. However, because users are more likely to post multiple sex-related tweets, the performance is much higher in practice. After accounting for the prediction errors, we identified 2,166 users who post sex-related content on Twitter, and are likely Black or Hispanic men living in Los Angeles from an initial set of 47.4 million tweets collected over 8 months. Despite the imprecision of the current automatic detection methods, our results suggest that it is possible to find men in the target age bracket living in the right location (Los Angeles) and characterize their language on Twitter to a degree that makes it feasible to then contact them for enrollment into epidemiological studies without reaching out to too many users that do not fit the demographic profile. Contacting them will allow us to validate the automatic methods, and complete the missing demographic information, if any.

The classifier has the potential to identify significant numbers of Black and Hispanic men at high HIV risk inclusive of all sexual identities. We are able to systematically analyze large quantities of social media data from men, a volume of data that would be infeasible to manually review. This approach is a significant departure from current online recruitment procedures for HIV prevention and care efforts, which are often limited by the under recruitment of men of color and men who have sex with women[1].

Funding

This work was supported by National Institute on Drug Abuse (NIDA) grant number R21DA049572-01 to SR. The content is solely the responsibility of the authors and does not necessarily represent the official view of NIDA.

References

[1] José A Bauermeister, Marc A Zimmerman, Michelle M Johns, Pietreck Glowacki, Sarah Stoddard, and Erik Volz. Innovative recruitment using online networks: lessons learned from an online study of alcohol and other drug use utilizing a web-based, respondent-driven sampling (webrds) strategy. *Journal of Studies on Alcohol and Drugs*, 73(5):834–838, 2012.

[2] Danielle E Ramo, Theresa MS Rodriguez, Kathryn Chavez, Markus J Sommer, and Judith J Prochaska. Facebook recruitment of young adult smokers for a cessation trial: methods, metrics, and lessons learned. *Internet Interventions*, 1(2):58–64, 2014.

[3] Jane Topolovec-Vranic and Karthik Natarajan. The use of social media in recruitment for medical research studies: a scoping review. *Journal of medical Internet research*, 18(11):e286, 2016.

[4] Annice E Kim, Timothy Hopper, Sean Simpson, James Nonnemaker, Alicea J Lieberman, Heather Hansen, Jamie Guillory, and Lauren Porter. Using twitter data to gain insights into e-cigarette marketing and locations of use: an infoveillance study. *Journal of medical Internet research*, 17(11):e251, 2015.

[5] Abeed Sarker, Graciela Gonzalez-Hernandez, Yucheng Ruan, and Jeanmarie Perrone. Machine learning and natural language processing for geolocation-centric monitoring and characterization of opioid-related social media chatter. *JAMA network open*, 2(11):e1914672–e1914672, 2019.

[6] Abeed Sarker, Graciela Gonzalez-Hernandez, and Jeanmarie Perrone. Towards automating location-specific opioid toxicosurveillance from twitter via data science methods. *Studies in health technology and informatics*, 264:333, 2019.

[7] Mark Dredze, Michael J Paul, Shane Bergsma, and Hieu Tran. Carmen: A twitter geolocation system with applications to public health. In *AAAI workshop on expanding the boundaries of health informatics using AI (HIAI)*, volume 23, page 45. Citeseer, 2013.

[8] Alastair van Heerden and Sean Young. Use of social media big data as a novel hiv surveillance tool in south africa. *Plos one*, 15(10):e0239304, 2020.

[9] Sean D Young, Caitlin Rivers, and Bryan Lewis. Methods of using real-time social media technologies for detection and remote monitoring of hiv outcomes. *Preventive medicine*, 63:112–115, 2014.

[10] Robin Stevens, Stephen Bonett, Jacqueline Bannon, Deepti Chittamuru, Barry Slaff, Safa K Browne, Sarah Huang, and José A Bauermeister. Association between hiv-related tweets and hiv incidence in the united states: Infodemiology study. *Journal of Medical Internet Research*, 22(6):e17196, 2020.

[11] Sean S. Barnes, Viren Kaul, and Sapna R. Kudchadkar. Social media engagement and the critical care medicine community. *Journal of Intensive Care Medicine*, 34(3):175–182, 2019.

[12] Matthias Briel, Kelechi Kalu Olu, Erik Von Elm, Benjamin Kasenda, Reem Alturki, Arnav Agarwal, Neera Bhatnagar, and Stefan Schandelmaier. A systematic review of discontinued trials suggested that most reasons for recruitment failure were preventable. *J Clin Epidemiol*, 80:8–15, 2016.

[13] Michael J. Wilkerson, Jared E. Shenk, Jeremy A. Grey, Simon B.R. Rosser, and Syed W. Noor. Recruitment strategies of methamphetamine-using men who have sex with men into an online survey. *J Subst Use*, 20(1):33–37, 2015.

[14] Jane Topolovec-Vranic and Karthik Natarajan. The use of social media in recruitment for medical research studies: A scoping review. *J Med Internet*, 18(11):e286, 2016.

[15] Hamza Shaban. Twitter reveals its daily active user numbers for the first time, 2019.

[16] Lauren Sinnenberg, Alison M. Buttenheim, Kevin Padrez, Christina Mancheno, Lyle Ungar, and Raina M. Merchant. Twitter as a tool for health research: A systematic review. *American journal of public health*, 107(1):e1–e8, 2017.

[17] Nicholas J. Hulbert-Williams, Rosina Pendrous, Lee Hulbert-Williams, and Brooke Swash. Recruiting cancer survivors into research studies using online methods: a secondary analysis from an international cancer survivorship cohort study. *ecancer*, 13(990), 2019.

[18] David Cavallo, Rock Lim, Karen Ishler, Maria Pagano, Rachel Perovsek, Elizabeth Albert, Sarah Koopman Gonzalez, Erika Trapl, and Susan Flocke. Effectiveness of social media approaches to recruiting young adult cigarillo smokers: Cross-sectional study. *J Med Internet Res*, 22(7):e12619, 2020.

[19] Tiernan J. Cahill, Blake Wertz, Qiankun Zhong, Andrew Parlato, John Donegan, Rebecca Forman, Supriya Manot, Tianyi Wu, Yazhu Xu, James J. Cummings, Tricia Norkunas Cunningham, and Catharine Wang. The search for consumers of web-based raw dna interpretation services: Using social media to target hard-to-reach populations. *J Med Internet Res*, 21(7):e12980, 2019.

[20] Christopher Whitaker and Fear Nicola Stevelink, Sharon. The use of facebook in recruiting participants for health research purposes: A systematic review. *J Med Internet Res*, 19(8):e290, 2017.

[21] Evan A. Krueger and Sean D. Young. Twitter: A novel tool for studying the health and social needs of transgender communities. *JMIR Ment Health*, 2(2):e16, 2015.

[22] Katja Reuter, Praveen Angyan, NamQuyen Le, Alicia MacLennan, Sarah Cole, Ricky N. Bluthenthal, Christianne J. Lane, Anthony B. El-Khoueiry, and Thomas A. Buchanan. Monitoring twitter conversations for targeted recruitment in cancer trials in los angeles county: Protocol for a mixed-methods pilot study. *JMIR Res Protoc*, 7(9):e177, 2018.

[23] Katja Reuter, Praveen Angyan, NamQuyen Le, and Thomas A. Buchanan. Using patient-generated health data from twitter to identify, engage, and recruit cancer survivors in clinical trials in los angeles county: Evaluation of a feasibility study. *JMIR Preprints*.

[24] Su Golder, Stephanie Chiuve, Davy Weissenbacher, Ari Klein, Karen O'Connor, Martin Bland, Murray Malin, Mondira Bhattacharya, Linda J. Scarazzini, and Graciela Gonzalez-Hernandez. Pharmacoepidemiologic evaluation of birth defects from health-related postings in social media during pregnancy. *Drug Saf.*, 42(3):389–400, 2019.

[25] Los Angeles Times. Mapping l.a. neighborhoods.

[26] Rebecca Knowles, Josh Carroll, and Mark Dredze. Demographer: Extremely simple name demographics. In *EMNLP Workshop on Natural Language Processing and Computational Social Science*, 2016.

[27] Zijian Wang, Scott Hale, David Ifeoluwa Adelani, Przemyslaw Grabowicz, Timo Hartman, Fabian Flöck, and David Jurgens. Demographic inference and representative population estimates from multilingual social media data. In *The World Wide Web Conference*, pages 2056–2067, 2019.

[28] Francisco M R Pardo, Paolo Rosso, Manuel Montes-y Gómez, Martin Potthast, and Benno Stein. Overview of the 6th author profiling task at pan 2018: Multimodal gender identification in twitter. In *CLEF*, 2018.

[29] Su Golder, Robin Stevens, Karen O'Connor, Richard James, and Graciela Gonzalez-Hernandez. Who is tweeting? a scoping review of methods to establish race and ethnicity from twitter datasets. *SocArXiv*, 2021.

[30] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[31] Robert J Stoller. *Sex and gender: The development of masculinity and femininity*. Routledge, 2020.