Using Noisy Self-Reports to Predict Twitter User Demographics

Anonymous Author(s)

ABSTRACT

Computational social science studies often contextualize content analysis within standard demographics. Since demographic attributes are unavailable on many social media platforms, such as Twitter, numerous studies have inferred demographic traits automatically. Despite many studies presenting proof of concept inference of race and ethnicity, training of practical systems remains elusive since there are few annotated datasets. Existing datasets are small, errorful, or fail to cover the four most common racial and ethnic groups in the United States. We present a method to identify self-reports of race and ethnicity from Twitter profile descriptions. Despite errors inherent in automated supervision, we train models sufficiently accurate to identify demographics when measured on a gold standard self-report survey. The result is a reproducible method for creating large-scale training resources for race and ethnicity.

ACM Reference Format:
Anonymous Author(s). 2019. Using Noisy Self-Reports to Predict Twitter User Demographics. In *Proceedings of ACM Conference (The Web Conference 2020).* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Contextualization of population studies with demographic characteristics forms a central method of analysis within the social sciences. Standard demographic panels included in telephone surveys across political science, public health and other domains enable the sub-population analysis of opinions and trends. Demographic attributes such as age, gender, race/ethnicity and location often serve as proxies for important socio-cultural groups. As the social sciences increasingly rely on computational analyses of online text data, limitations imposed by a lack of availability of demographic attributes hinder comparison of these studies to traditional methods.

Computational social science increasingly utilizes methods for the automatic inference of demographic attributes from social media, such as Twitter [10, 54]. Demographic attributes have been included in social media studies in varied domains, such as health [16, 23], politics [49], and linguistics [27]. Off-the-shelf software packages support the inference of gender [36] and location [24, 58].

Unlike age or geolocation, race and ethnicity are sociocultural categories with competing definitions and measurement approaches [15, 18, 65]. Despite this complexity, understanding race and ethnicity is crucial for public health research [25]. For example, analyses that explore Twitter users' discussions of mental health [42] must be able to consider racial disparities in healthcare [1, 61] or online interactions [11, 21]. Despite the importance of race and ethnicity in these studies, and multiple proof-of-concept classification studies, there are no readily-available systems that can cover the major racial/ethnic groups in the United States. This gap is primarily because all publicly-available data resources have major limitations.

What challenges prevent the development of a large, high-quality dataset for training a race/ethnicity inference system for social media? First, the dataset should include the most common categories to match standard demographics panels. In this paper, we focus on the most prominent categories present in the United States. Second, the dataset must be sufficiently large to support training accurate systems. Third, the dataset should be reproducible. On Twitter, all datasets shrink over time as users delete or make private their accounts. On any social media platform, domain drift will make learned features less useful over time [33].

We present a method for automatically constructing a large dataset for race and ethnicity. We initially use keyword-matching to construct a large corpus of tweets to find Twitter users who self-identify with a racial or ethnic group, building on past work that explored considered Twitter self-reports [4, 16]. We then learn a set of filters to remove users who match keywords but do not actually self-report their demographics. Our approach can be easily replicated in the future, constructing an updated dataset for training. Our use of keywords for automatic supervision is inherently noisy – self-descriptions can be hard to automatically interpret – but in aggregate our large dataset enables classification results that greatly outperform previous small, human-labeled datasets. We demonstrate the quality of a classifier trained our data by evaluating on a gold-standard survey dataset of self-reported labels [56].

2 ETHICAL CONCERNS AND CONSIDERATIONS

Complexities of racial identity raise important ethical considerations, requiring discussion of the benefits and harms of this work [6].

The benefits of such research are clear. Consider public health, which has the goal of preventing disease and improving the overall health of the population. Numerous studies have used Twitter and other social media data to derive insights on health behaviors and to create health-based interventions [50, 51, 62], and these methods have transformed whole areas of public health research which previously lacked in accessible data [3]. Demographic inference is key for many of these studies, as it enables alignments between social media derived insights and data from more traditional sources.

The concerns and potential harms are more complex. The use of Twitter data for research entails complex issues of privacy and informed consent of study participants [28, 43]. While Twitter's privacy policy states that the company "make[s] public data on Twitter available to the world," many users may not be aware of the scope or nature of research conducted using their data [45]. As participant consent must be *informed* to be valid, we should have increased concerns about the knowledge gaps between user

The Web Conference 2020, April 2020, Taipei, Taiwan

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Citation	% Missing	# Users	% W	% B	% H/L	% A
Preoţiuc-Pietro et al. [57]	4.7	3572	80.8	9.5	6.1	3.6
Culotta et al. [19]	60.0	308	50.0	19.5	30.5	0
Volkova and Bachrach [66]	36.5	3174	48.0	35.8	8.9	3.0
Total Matching Users	-	2.50M	26.8	53.8	11.3	8.1
Group-Person	-	122k	50.5	40.5	1.6	7.4
Weighted-Group	-	228k	40.1	47.4	5.7	6.9
Balanced-Group-Person	-	31k	25.0	25.0	25.0	25.0

Table 1: Top: Previously-published Twitter datasets annotated for race/ethnicity. Bottom: datasets collected in this work. "% Missing" shows the percent of users that could not be scraped in 2019, and "# Users" shows the number users that are currently available. The categorical breakdown (White, Black, Hispanic/Latinx, Asian) is based on non-missing data.

understanding of the terms of service when conducting research on possibly sensitive issues such as race.

Specifically regarding demographic prediction, the same tools that can provide health insights can be used to track or intimidate minority groups. Recent work has show that women on Twitter, especially journalists and politicians, receive disproportionate amounts of abuse [22]. On Facebook, advertisers have used the platform's knowledge of users' racial identities to illegally discriminate when posting job or housing ads [2, 5]. This abuse and discrimination is already widespread, even without off-the-shelf tools for predicting user demographics. Twitter's developer terms of service prohibit using Twitter content to "target, segment, or profile individuals" based on several sensitive categories, including racial or ethnic origin.[1] However, demographic inference tools also have a role in *preventing* such abuses, for example, by enabling studies on abusive behavior and hate speech on social media.

Another concern of any predictive model for sensitive traits is that a descriptive model could be interpreted as a prescriptive assessment. Errors made on individual users could subject those users to possible harm from downstream applications reliant on demographic inference models or by supporting a reductive or essentialist framework for racial or gender identity [32].

On the whole, we believe these tools provide significant benefits that justify the potential risks in their development. However, these tools should only be used for population-level analyses and **not** the identification and analysis of individual users. Even with high error rates at an single-user level, aggregating predictions across a population can provide valuable insights. We note that aggregate analyses of content is explicitly supported by Twitter's restricted uses of APIs.[2] Future applications must reconsider cost-benefit trade-offs as technologies and environments change.

We make our data available to other researchers, but with limitations on its use. Specifically, we require that researchers obtain approval by an Institutional Review Board (IRB) or similar ethics committee before obtaining our data. We explicitly exclude certain applications, such as targeting of individuals based on race or ethnicity. Finally, our analysis of social media for public health research has been reviewed and deemed exempt (45 CFR 46.101(b)(4)) by our IRB.

3 DATASETS FOR RACE AND ETHNICITY

Historically in the United States, recognized racial categories have varied over time [31, 40], with the current census – and many surveys – recording self-reported racial categories as White, Black, American Indian, Asian, and Pacific Islander. In the U.S., questions of ethnicity often only ask regarding Hispanic or Latinx origin; however, there is not necessarily a clear distinction between race and ethnicity [12, 17, 30]. Individuals may identify as both a race and an ethnicity, and 2% of Americans identify as multi-racial [35]. Because of the limited availability of data resources, we only consider the four largest race/ethnicity groups, which we model as mutually exclusive: White, Black, Asian, and Hispanic/Latinx. Our methodology is quite flexible to a more comprehensive choice of demographic labels, but we do not yet have the means to validate more fine-grained or intersectional approaches.

The top of Table 1 lists previously published datasets for race/ethnicity. While each dataset has been used for training computation models, each has drawbacks that limit their value for downstream modeling applications. A persistent issue is that since only userids can be shared, user account deletions over time cause substantial missing data (e.g. no Asian users remain for Culotta et al. [19]).

There have been a variety of approaches used to create annotated datasets for demographics. Culotta et al. [19] and Volkova and Bachrach [66] relied on manual annotation, noting inter-annotator agreement estimated at 80% and Cohen's κ of 0.71, respectively. Preoţiuc-Pietro et al. [57] conducted a survey to collect self-reported demographics. Pennacchiotti and Popescu [54] automatically label African American users using profile mentions. Manual annotation assumes that racial identity can be accurately perceived by others, an assumption that has serious flaws for gender and age [29, 55]. Rule-based or statistical systems for data collection can be effective [10, 13], but raise concerns about selection bias: if we only label users who take a certain action, a model trained on those users may not generalize to users who do not take that action [69]. Explicitly querying for demographics in a survey provides a gold-standard, but yields small or skewed datasets due to survey expense [56].

We take a hybrid approach in this paper, relying on automated labeling based on racial self-identification and using minimal manual labeling to refine our dataset labels. We present our method as an automated pipeline so that it can be repeated in the future to update our dataset. We evaluate our label quality via an experimental evaluation on self-reported attributes [56].

[1] https://developer.twitter.com/en/developer-terms/agreement

[2] https://developer.twitter.com/en/developer-terms/more-on-restricted-use-cases.html

	Raw	Color	Plural	Bigram	Quote	All
Precision	76.74	78.57	76.74	82.50	78.57	**86.84**
Users removed by filter(s)	-	314k	212k	281k	4k	784k

Table 2: Applying our four `WG` filters (§ 4) individually and altogether. Precision is calculated on our manually-annotated dev set from Appendix B.1.

4 DATA COLLECTION OF SELF-REPORTS

We begin our data collection of racial self-reports by constructing a regular expression for terms associated with racial identity. We scrape all tweets with a user description that matches our query regex:

```
(black|white|caucasian|asian|hispanic|latin[oax]?)
```

We apply this filter to tweets collected from Twitter's public sample streaming API (1% of tweets) from July 2011 to July 2019, producing an initial dataset of 88GB (compressed). For users who appear multiple times in the streaming API, we consider their latest descriptions. We process matching tweets by recursively searching the `retweeted_status` and `quoted_status` fields to extract additional tweets to match with the regular expression.

Our initial user extraction relies on exact matches (using word boundaries) for our We begin by counting users with an exact match (with word boundaries) of a query word to establish an upper bound of the total possible users who self-report with our identified terms. This heavily skews towards 'white' and 'black', since those are colors used in many contexts. This produces 2.67M matching users, of which 2.50M match exactly one racial/ethnic category. The distribution of these categories are shown in the 'Total Matching Users' row in the bottom of Table 1, which also contains statistics for the following three datasets we consider.

Group-Person (GP). We use a regex that matches a query word followed by (and separated only by whitespace from) one of the following self-report words: `man, woman, person, individual, guy, gal, boy, or girl`, e.g. "Black woman" or "Asian guy." This approach exacerbates the skewness of the previous data, as 91.0% of the resulting 122k users are labeled as either white or black.

Weighted-Group (WG). While the Group-Person dataset uses a more restrictive regex, we now consider filtering users' descriptions to more accurately identify self-reports. We learn a 'self-report' score which gives high scores to descriptions which are more likely to be self-reporting race/ethnicity, and low scores to user descriptions which match our initial keywords but are not self-reporting. We learn this score by leveraging lexical co-occurrence, an important cue for word associations [14, 63]. Our score combining the relative frequencies of co-occurring words within a fixed window, which we down-weigh by the distance between query and co-occurring self-report words. The intuition here is that if we believe "farmer" is a valuable self-report word such that "Black farmer" is a high-scoring self-report, then "Black beans farmer" should have a lower score due to the distance between the query term and the self-report word. We consider two approaches for this weighting, one which leverages the concept of TF-IDF scoring of query terms, and another which does not [59]. These scoring details are in Appendix B. We use a small manually-labeled tuning set to threshold this score, described in Appendix B.1.

We found four techniques helpful in improving our `WG` dataset. First, many non-self-report descriptions matched "black" and "white" in addition to other colors, so we filtered out all words from a color-list [7]. Second, we used NLTK `TweetTokenizer` [8] to obtain part-of-speech tags, and found that an adjective query word preceding plural nouns were unlikely to be self-reports (e.g. "white people"). Third, we curate a list of word bigrams that most frequently contain a query but are unlikely to be self-reports (e.g. "black sheep").[3] This produces an intersection of 286 non-self-report bigrams, which we filter out. Finally, query words that appear inside quotation marks are ignored. Table 2 shows how these methods decrease the size of our datasets but increase precision on our tuning set.

Balanced Group-Person (BGP). While the Weighted-Group dataset contains more users than the Group-Person Dataset, both are quite class-imbalanced. To build a dataset with equal representation across all four groups, we start with our smallest group and extract all 7756 Hispanic/Latinx-labeled users. Then, using our self-report scores, we take the 7756 highest-scoring users from each of the other three categories and add them to our dataset. Thus, our balanced dataset includes all Asian-labeled users from the Group-Person dataset and most from the Weighted-Group dataset, only a fraction of the White and Black-labeled users from the Group-Person dataset and no users from the Weighted-Group dataset.

We highlight these datasets and their label distributions in the bottom half of Table 1 and refer to them by their acronyms (`GP, WG, BGP`) in the experimental results in Table 3.

5 EXPERIMENTAL EVALUATION

Given our use of automated methods to label the collected datasets, we need an independent approach to validate the quality of our labels. For each collected dataset, we train a demographic classifier on our data and then use the trained model to predict labels for the gold-standard data [57]. If models trained on our data accurately classify self-reported demographics, then our data is valuable for downstream applications.

The evaluation test set is the gold-standard dataset [57]. While this dataset has the highest-quality labels of any published work on social media demographics, due to the high cost of surveys, the dataset only contains 4.1k users, of which we can collect data on only 3.6k. We use other previously-published datasets to produce a development set and a baseline dataset. Both the dataset from Culotta et al. [19] and that of Volkova and Bachrach [66] used manual crowdsourced annotations to label Twitter users. We combine these to produce a dataset of 3.5k users, which we randomly split into a 60% training and 40% dev set. Our experimental evaluation compares models trained on this crowdsourced training set, models trained on our self-report data alone, and models trained on both. We use the dev set to perform model selection in all experiments.

As the test set has an extremely imbalanced distribution over its four racial/ethnic categories (see Table 1), we build sub-sampled dev and test sets that contain an equal proportion of each demographic group. The balanced dev set contains only 168 users; the balanced

[3] Frequencies are from the Google N-gram corpus [44]. We only use bigrams in the form of query + word with more than 100k occurrences.

	Names		Unigrams	
Dataset/Baseline	F1	Acc %	F1	Acc %
Random baseline	.250	25.0	.250	25.0
Majority baseline	.224	**80.8**	.224	80.8
Crowdsourced	.264	78.4	.372	83.5
GP	.325	80.5	.352	84.2
Crowdsourced+GP	.314	79.9	.368	83.9
WG	.323	78.7	.357	**84.3**
Crowdsourced+WG	.233	78.1	.345	83.9
BGP	**.339**	65.9	.396	82.9
Crowdsourced+BGP	.319	57.8	**.422**	84.0

(a) Imbalanced prediction task for both single-tweet name and many-tweet unigram models.

	Names		Unigrams	
Dataset/Baseline	F1	Acc %	F1	Acc %
Random baseline	.250	25.0	.250	25.0
Majority baseline	.100	25.0	.100	25.0
Crowdsourced	.177	28.6	.252	36.3
GP	.282	33.3	.277	39.4
Crowdsourced+GP	.293	34.8	.245	33.4
WG	.274	35.5	.262	35.6
Crowdsourced+WG	.139	25.1	.260	38.9
BGP	**.367**	36.8	.409	45.1
Crowdsourced+BGP	.360	**37.0**	**.424**	**46.9**

(b) Balanced prediction task for both single-tweet name and many-tweet unigram models.

Table 3: Experimental results for models trained on the crowdsourced datasets and our self-report datasets. The best result in each column is in bold. Dataset abbreviations are defined in § 4. '+' indicates a combined dataset of crowdsourced data plus our self-report data. Section 5 and Appendix C contain the training and evaluation details.

test set contains only 452 users. Tables 3a and 3b shows imbalanced results and balanced results, respectively.

To further highlight the differences between the imbalanced and balanced datasets, we evaluate on both total accuracy and macro-averaged F1. We also show the performance of two naïve strategies: randomly guessing across the four demographic categories, and deterministically guessing the majority category. Because of the class imbalance, 'Majority Baseline' strategy will achieve 80.8% imbalanced accuracy. However, it only achieves an imbalanced F1 score of .224.

Finally, we include a method of adapting across disparities in the training versus test dataset. After training our models, we use the dev set to estimate the per-class precision and recall of our classifier, and then use a greedy prediction-rebalancing algorithm to match our test-time prediction to a prior over the test set's class distribution. The full details of this rebalancing approach and its effects on our classification accuracy are in Appendix D.

We stress these evaluation details because extreme class-imbalance can cause serious complications. Models trained to do well on the majority class at the expense of minority classes could bias downstream analyses by under-representing minority groups. In health, which has well known disparities between majority and minority groups [39], this could produce research results that exacerbate rather than ameliorate inequalities.

5.1 Demographic Prediction Models

We consider two demographic inference models, which we train on each training set and evaluate in both imbalanced and balanced experimental settings. First, we consider a system that only has access to a single tweet per user, which severely reduces its feature set, but can be run over a large Twitter corpus without requiring an extensive tweet history from each user. We use the model of Wood-Doughty et al. [68], which learns a character-based convolutional neural network (CNN) of the user's name as well as features from the user metadata, such as the user's verification status and ratio of followers to friends. These features are passed through a two-layer MLP to produce a distribution over the label classes. This model is referred to as 'Names' in Table 3.

Second, while the name-based model requires only a single tweet per user, models that rely on content tend to perform better [66, 68]. Therefore, we consider a content-based method that examines aggregated features across many tweets for a user. We use the model of Volkova and Bachrach [66] who include unigrams from a user's tweet history in a logistic regression classifier. We include as features the 77k non-stopword unigrams that occur at least twice in the development set. We (attempted to) download the 200 most recent tweets for each user from the Twitter API. The total number of users for which we could download data is reflected in Table 1; users for which API queries failed count towards the '% Missing' column. Further data collection details are in Appendix C. This model is referred to as 'Unigrams' in Table 3.

6 EXPERIMENTAL RESULTS AND DISCUSSION

Table 3 shows the results for each model and dataset. The many-tweet Unigrams model outperformed the single-tweet Names model in both F1 and accuracy across multiple datasets. This is consistent with past published work; the Unigrams model has much more data per user than the Names model.

On the balanced evaluation sets, models trained on our collected datasets obtain improvements of up to 10.6% accuracy and .172 F1 over models trained only on the previously-published datasets.

In the imbalanced evaluations, we see a large trade-off between accuracy and F1, as the models can achieve better overall accuracy when they learn to ignore the Asian and Hispanic/Latinx classes. In the imbalanced setting, the trivial 'Majority Baseline' strategy achieves a better accuracy than any of the Name-based models we trained, yet still has the worst F1 score. Our BGP dataset, which is explicitly designed to learn a model that does not ignore any of the four demographic labels, achieves the best F1 scores in both the imbalanced and balanced evaluations. In the balanced evaluations where overall accuracy cannot be improved by ignoring infrequent labels, the BGP models also achieve the best accuracy. In many cases,

Asian	Black	Hispanic/Latinx	White
love	np	love	mytwitteranniversary
asian	soundcloud	tbt	blessed
mytwitteranniversary	tbt	music	soundcloud
tbt	love	beliebers	tbt
gameofthrones	mytwitteranniversary	mytwitteranniversary	love
food	blessed	family	gameofthrones
giveaway	music	believe	newprofilepic
selfie	gameofthrones	nyc	np
lol	nowplaying	mtvhottest	sorrynotsorry
bts	blackgirlmagic	soundcloud	loveisland

(a) Top 10 hashtags by the number of users who tweet about it for each racial group. Note the hashtags are lowercased to combine ones with the same topic.

Asian	Black	Hispanic/Latinx	White
liked	avrillavigne	justinbieber	bc
visit	ni##as	justin	realdonaldtrump
hahaha	black	online	snapchat
art	ni##a	follow	dog
youtube	wit	unfollower	holy
lunch	dat	unfollowers	drunk
found	sis	stats	dad
haha	libra	follower	pizza
hi	da	followers	cat
sexy	capricorn	la	f##king

(b) Top 10 keywords via lexical variation computed by SAGE. Inserted # partially mask explicit language.

Table 4: Top hashtags and keywords per group.

a combination of our self-report datasets and the crowdsourced datasets does better than our self-report dataset alone, but this is not true across all evaluations. Further ablation studies could provide more explanation as to how features learned from our data differ from features learned from the crowdsourced datasets.

These empirical results demonstrate that incorporating our self-report datasets into a training set produces better demographic inference models on held-out, gold-standard labels. However, any downstream application to analyzing real-world demographic trends in aggregate may require considering problem-specific trade-offs. If a researcher wants to study health behaviors within a specific racial group, they may be willing to compromise classification accuracy in all but that specific group.

In many health surveillance tasks, a researcher may want to know if a health behavior or exposure to a risk factor varies significantly between two groups [37]. If we want to conduct such an analysis on Twitter using a demographic classifier such as the ones trained in this work, the accuracy of a trained classifier may have a direct link to what magnitude of health disparities can be detected. Such research requires a careful contextualization of what conclusions can be drawn from the available data and models; differences between groups may be under- or over-exaggerated by classifier errors.

7 DIFFERENCES IN TWITTER BEHAVIORS ACROSS RACIAL GROUPS

Our experimental results demonstrate that our noisy self-report data, in aggregate, offers better predictive power than a smaller dataset of human-labeled data. However, these classification results do not provide insights about the users we collect who self-report their demographic categories. Are these users representative of the typical Twitter user? Or is our dataset collection technique producing a biased sample of users, and if so how? A particular challenge with our work is that our initial collections of keyword matching yielded a dataset that is heavily skewed towards the 'White' and 'Black' keywords. Are users' propensity to use these keywords in their profile descriptions were completely independent of the other ways in which they use Twitter?

We explore these questions using a variety of quantitative analyses of linguistic and platform-specific behaviors. There are two different interpretations to keep in mind when considering these group-level differences. On the one hand, the Twitter user behaviors we measure may correlate with demographic categories [69]. However, it may also be that the *behavior of self-reporting* correlates with these

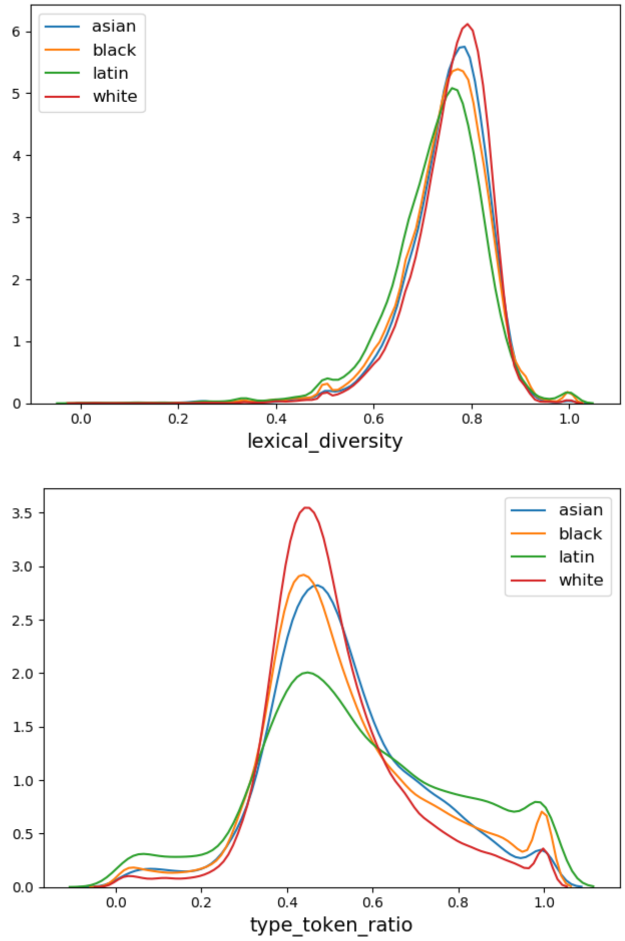


Figure 1: Plots of the Lexical diversity and TTR scores within each group. Lexical diversity is quite similar across groups, whereas TTR has larger divergences between groups.

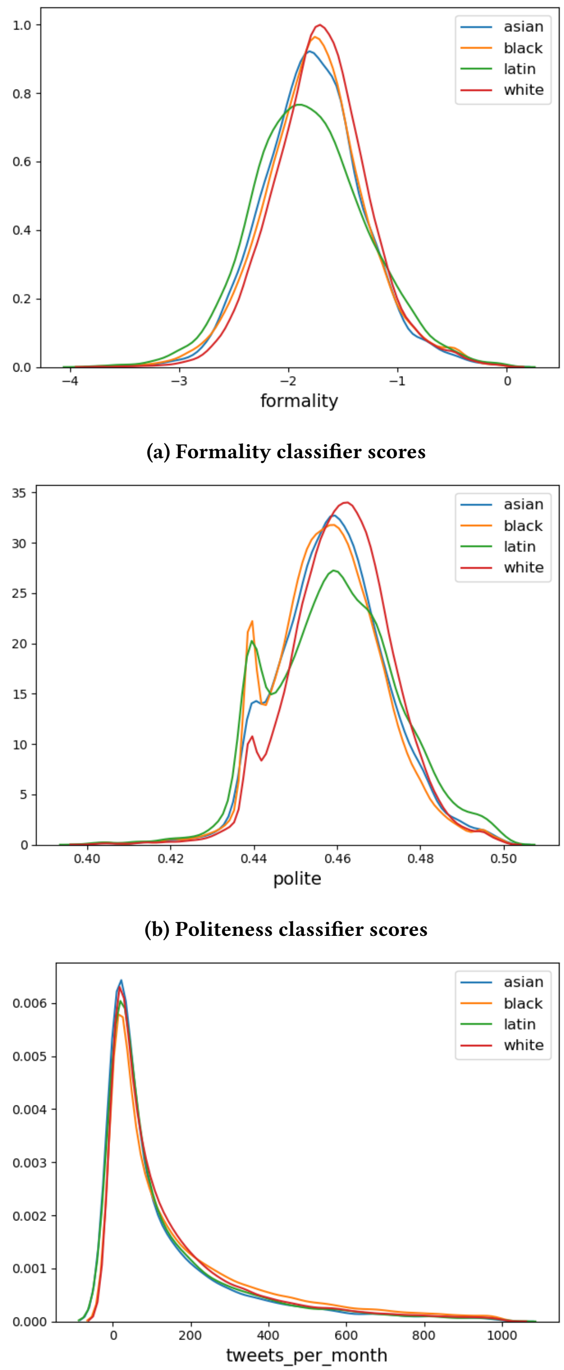
behaviors. These are not mutually-exclusive, and our current methods cannot distinguish between group differences themselves or selection bias that may depend on racial category or Twitter usage behaviors.

List-based Features. We first seek to understand possible topical differences between groups in our collected datasets, by compiling a ranked list of the most popular hashtags per group. We then take a similar approach and compile ranked lists of popular emojis, emoticons, and part-of-speech tags. We also use a SAGE [26] lexical variation implementation to find the words that most distinguish each demographic group, following [46]. We treat each of the lists as a categorization of each group. To compare across groups, we look at the top k items in each list and calculate Kendall τ rank correlation coefficients for each pair of demographic groups [47]. These coefficients vary between -1 for a perfect negative correlation and 1 for a perfect positive correlation. For emojis and emoticons, all correlations are negative for smaller k values, but they tend in a positive direction as we increase k. Despite this general trend, we see strong differences in content usage between groups. For hashtags, in particular, correlations are strongly negative for all values of k, suggesting that groups labeled by our method substantially differ in the topics they discuss. As a qualitative look at topical differences, we show the top-10 hashtags and SAGE keywords for each group in Table 4. The table of all pairwise Kendall τ calculations is in Table 8 in the Appendix.

We see that some hashtags, e.g. `#MyTwitterAnniversary` and `#GameOfThrones` are popular across all the labeled groups, yet many hashtags only appear in the top 10 for a single group. A more thorough analysis could conduct such an analysis periodically, to explore whether universally-popular hashtags are popular with each group simultaneously. While our initial data collection keywords are all in English, we do not make an attempt to limit our analysis to English-language speakers. Thus, differences in topic discussion may be confounded by users' native language(s).

Quantitative Linguistic Features. Lexical features are widely used for classifying users on Twitter [9, 54, 60]. To understand possible linguistic differences between collected groups, we follow §3.1 of Inuwa-Dutse et al. [34] and for each user in our group, we calculate Type-Token Ratio (TTR), Lexical Diversity [64], and the proportion of English contractions they use. TTR is defined as the number of unique tokens in a tweet divided by the total number of tokens in the tweet. We compute lexical diversity as the total number of tokens in a tweet without URLs, user mentions and stopwords divided by the total number of tokens in the tweet. A comparison of the mean of each quantitative linguistic features are in the Appendix in Table 7.

To explore difference between groups using previously-trained linguistic models, we consider quantitative models for evaluating formality [53] and politeness [20] of online text. The formality score is estimated with a regression model over lexical and syntactic features including n-grams, dependency parse, and word embeddings. For both models, we compute the average score over all the text from one user, and average over all the users within certain demographics group. We use the implementation released by the authors.[4] The linguistically-informed politeness classifier considers unigram features, as well as multiple politeness strategies including lexicons for



(a) Formality classifier scores

(b) Politeness classifier scores

(c) Average tweets per month

Figure 2: Plots of the formality [53] and politeness [20] scores within each group, as well as the average tweets per month. X-axis is truncated to highlight differences, but all three plots contain more than 95% of their respective distribution density.

[4] https://github.com/YahooArchive/formality-classifier

	Asian	Black	Hispanic/Latinx	White
% users in dataset	6.71	49.44	5.83	38.02
(m) % tweets from Android sources	$24.37^{*\dagger}$	$25.34^{*\ddagger}$	$23.96^{\dagger\ddagger}$	16.57
(m) % tweets from iPhone sources	46.15	47.56	40.88	64.98
(m) % tweets from iPad sources	1.46	1.07	1.40^{*}	1.29^{*}
(m) % tweets from desktop web	10.23	7.29^{*}	13.21	6.09^{*}
% users with 1+ tweets from Android	$38.95^{*\dagger}$	38.33^{*}	$39.41^{\dagger}$	25.46
% users with 1+ tweets from iPhone	60.28	58.21	54.89	75.37
% users with 1+ tweets from Desktop	43.34	30.59	44.87	31.04
% users with profile URL	34.09^{*}	29.71	34.75^{*}	24.78
% users with custom profile image	98.83	$99.29^{*\dagger}$	$99.24^{*\ddagger}$	$99.33^{\dagger\ddagger}$
% users with geotagging enabled	48.65^{*}	53.27	49.54^{*}	56.04
% users with 1+ geotagged tweet	8.35^{*}	6.46	7.81^{*}	5.43
Average statuses count	11974	18709	12449	14177
Average tweets per month	177.83	255.41	182.13	200.85
(m) % tweets that mention a user	59.73	58.71	60.44^{*}	61.77^{*}
(m) % tweets that include an image	20.44^{*}	17.20	18.39	19.17^{*}
(m) % tweets that include a URL	20.99	21.64	24.01	17.22

Table 5: Profile Behavioral Features. (m) indicates that a percent or average was computed via micro-averaging across users' tweets; all others are macro-averaged across users. Almost all differences in a row are statistically significant from one another, according to a Mann-Whitney U Test. However, if two entries in the same row share a superscript symbol, they are not significantly different at a 0.05 confidence level.

gratitude and positive or negative sentiment. We use the implementation released by the authors.[5] Plots of these metrics across each group are shown in Figures 2a and 2b. The plots of formality and politeness both appear roughly Normal except for a spike near 0.44 in the politeness plot. As the politeness classifier uses lexicon features, it may be that some of the spike is capturing a binary indicator of whether certain lexicon entries are present in a tweet. As with list-based features, these features may be heavily influenced by users' native language(s).

Profile Behavioral Features. Finally, we consider a few basic measures of Twitter usage, computed from the profile information of each user, following Wood-Doughty et al. [69]. Table 5 contains these the mean value of these features, describing broad range of basic user behaviors on the Twitter platform. We then plot the full distribution of average tweets per month for users in each group in Figure 2c. Almost all differences in these behavioral features are significant across groups. The biggest difference appears in device usage, where we see that White users are much more likely to have used an iPhone to tweet and much less likely to have used an Android to tweet, when compared against users of the other three demographic groups.

This table also provides interesting comparisons to prior work. Pavalanathan and Eisenstein [52] demonstrated that the use of Twitter geotagging was more prevalent in metropolitan areas and among younger users. Our construction of Table 5 follows past work that calculated similar profile behavior features for a random sample of 1M Twitter users in 2017 [69]. Comparing against those numbers, we see that across all our demographic groups, users in our datasets are much more likely to include a custom profile image or profile URL, or to enable geotagging on their profile.

Across all types of features we consider, we see many substantial differences between the different groups labeled by our data collection methods. This provides strong evidence that our data collection based on description keywords is correlated with actual underlying differences with how users in each group use the Twitter platform. However, it cannot reveal to us whether these differences are primarily correlated with racial/ethnic groups, or whether we see these differences primarily based on how users make the decision whether to self-report a race/ethnicity keyword.

8 LIMITATIONS AND FUTURE WORK

This work has presented a reproducible approach for automatically identifying self-reports of race and ethnicity to construct an annotated dataset for learning demographic inference models. While the automated annotations produced by our method are imperfect, we show that our data can supplement or replace manually-annotated data. This enables the development and distribution of tools to facilitate demographic contextualization in computational social science research.

There are several important extensions that should be considered. Our analysis only focuses on the United States; most countries have a unique cultural conceptualizations of race/ethnicity and unique demographic composition, and may require a country-specific focus. Our method covers four categories of race/ethnicity, it ignores smaller populations and multi-racial categories [35]. Additionally, we use a limited set of query terms, which ignores the diversity of how people may choose to self-report their identities. Finally, it is well known that there are biases in demographic inference [52, 69]. Aside from mislabeled users, selection bias and generalization are serious concerns. In future work, we strongly encourage the study of racial self-identities and social cultural issues as supported

[5] https://github.com/sudhof/politeness

by computational analyses. Furthermore, these issues should be considered with a global perspective, especially with regards to biases in our collection methods [38].

We will release our code and our annotated Twitter userids to enable comparison to our method, the training of new models, and for the construction of future updated datasets. Use of our data or models will require complying with a data use agreement and obtaining approval from an appropriate ethics board.

A PREPROCESSING, TOKENIZING, AND TAGGING

We lowercase all descriptions, and replace line break with full stop, as such symbols are commonly used to separate descriptors of themselves. We use NLTK Tweet Tokenizer [8] and its PoS tagging toolkit to tokenize the text and get the PoS tags. We collect a large dataset of candidate self-report words based on their co-occurrence with the regex and PoS tag pattern, `{I‘/I a}m (+ RB)( + DT) (+ JJ) + NN`, in a collection of 177M Twitter descriptions. We collect both adjectives and nouns from the pattern above. To reduce the impact of noisy PoS tagging, we then refine the candidate self-report words by keeping the words which are majorly tagged desirably for corresponding adjective and noun collections, based on Google N-gram tagging.

We filter out plural candidate self-report words using a PoS tag pattern, `JJ + NNPS/NNS`. This plural filter removes likely false-positives such as “white people.” We refer to the set of non-plural self-report words as S.

B CALCULATING THE ‘SELF-REPORT’ SCORE

The weighting strategies mentioned in § 4 are discussed in detail. The simple co-occurrence weighting is obtained by considering the occurrence $O_s(w_s)$ of self-report word w_s appears as self report and its occurrence $O(w_s)$ in general, denoted as

$$R = \sum_{w_s \in S^{win}} \frac{1}{D(w_s, q)} \cdot \frac{O_s(w_s)}{O(w_s)},$$

where S^{win} is the self-report words in the fixed window size, $D(w_s, q)$ denotes the distance between w_s and query word q.

The TFIDF weighting is computed as

$$R_{\text{tfidf}} = \sum_{w_s \in S^{win}} \frac{1}{D(w_s, q)} \cdot \frac{O_s(w_s)}{O(w_s)} \cdot \log \frac{\sum_{w \in S} O_s(w)}{O_s(w_s)}$$

B.1 Self-report Score Hyperparameters

To fine-tune our self-report score, we manually labeled a tuning set of 400 descriptions as to whether the user was self-reporting a matching query word. We discarded 6 that were organizations, and had an Krippendorff α 0.8058 within three annotators on the remaining 394. We adopted majority voting strategy for deciding the labels.

For the self-report score’s hyperparameters of window size, We select these hyperparameters based on the precision calculated on the development set. The performance of collected dataset based on different hyperparameter settings are shown in Table 6a and Table 6b, corresponding to the simple co-occurrence weighting and TFIDF weighting schema. We also considered increasing the exponent of the $1/D(w_s, q)$ term, but found it had no effect on precision.

To ensure that these chosen hyperparameters did not overfit to the tuning set, we sampled an additional 199 users for a test set. For those users, three authors manually annotated whether the Twitter users were actually self-reporting the racial demographics possibly implied by the query word matches. Using a three-label nominal scale of ‘yes,’ ‘no,’ or ’unsure,’ the three annotators achieved a Krippendorff’s alpha of 0.625. When we convert the manual

Window size Threshold	3	5	7	9	11	13
1e-5	50.64 / 1091k	49.51 / 1421k	49.08 / 1527k	49.11 / 1578k	48.90/ 1604k	48.91 / 1622k
1e-3	52.32 / 1069k	49.75 / 1397k	49.07 / 1499k	49.55 / 1553k	49.11 / 1580k	48.68 / 1600k
1e-1	59.57 / 613k	62.62 / 734k	61.82 / 793k	58.62 / 836k	58.97 / 864k	57.98 / 887k
3e-1	80.00 / 256k	78.72 / 286k	77.55 / 307k	77.55 / 322k	77.55 / 333k	76.00 / 341k
3.5e-1	85.29 / 205k	**86.84 / 227k**	82.50 / 244k	79.07 / 258k	79.55 / 267k	79.55 / 274k
4e-1	81.82 / 134k	83.33 / 152k	84.62 / 169k	78.57 / 180k	78.57 / 188k	79.31 / 194k
4.5e-1	100 / 64k	100 / 80k	100 / 96k	100 / 107k	100 / 114k	100 / 120k

(a) Selection of hyper-parameters for simple co-occurrence weighting based on the development set. The contents are in the form of precision on dev set / # of users in the dataset.

Window size Threshold	3	5	7	9	11	13
0.3	53.23 / 861k	53.33 / 1077k	51.83 / 1161k	51.45 / 1222k	51.98 / 1258k	51.10 / 1286k
0.7	56.00 / 666k	56.14 / 804k	53.72 / 869k	52.80 / 919k	53.54 / 950k	54.20 / 975k
1.3	65.71 / 467k	66.67 / 549k	65.06 / 596k	64.71 / 634k	61.54 / 659k	61.29 / 679k
1.7	73.81 / 329k	73.47 / 388k	66.67 / 427k	66.67 / 456k	67.19 / 477k	66.67 / 495k
2.3	85.00 / 134k	80.95 / 171k	78.26 / 220k	80.00 / 249k	76.92 / 268k	76.92 / 280k
2.5	86.67 / 103k	87.50 / 133k	88.24 / 164k	89.47 / 184k	**89.47 / 200k**	85.00 / 211k
2.7	81.82 / 79k	84.62 / 105k	85.71 / 130k	86.67 / 146k	86.67 / 158k	86.67 / 167k

(b) Selection of hyper-parameters for TFIDF weighting based on the development set. The contents are in the form of precision on dev set / # of users in the dataset.

Table 6: Hyper-parameter selection for simple and TFIDF weighting.

	# Users	Lexical Diversity	Contractions per tweet	TTR	Hashtags per tweet	Formality	Politeness
Asian	9442	0.75120	0.07544	0.53285	0.15543^*	-1.76991	0.45948
Black	70838	0.74655	0.06657	0.53228	$0.09560^\dagger$	-1.74977	0.45840
Hispanic/Latinx	8349	0.73061	0.05065	0.56258	0.14521^*	-1.80172	0.46091
White	57724	0.75913	0.08521	0.51000	$0.08132^\dagger$	-1.69685	0.46136

Table 7: Comparison of the mean values for each numerical feature between racial groups. Almost all differences are significant. Within each column, only numbers with a shared superscript symbol are not significantly different at a 0.05 **confidence level when using a Mann-Whitney U test.**

	Emojis			Emoticons			Hashtags			PoS bigrams			PoS trigrams		
Top k	20	50	80	20	50	80	20	50	80	20	50	80	20	50	80
A vs. B	-0.67	-0.26	-0.05	-0.19	0.10	0.19	-0.85	-0.87	-0.86	0.29	0.19	0.79	0.18	0.19	0.46
A vs. H/L	-0.10	-0.07	0.00	-0.02	0.08		-0.84	-0.86	-0.86	0.55	0.02	0.58	0.20	0.02	0.25
A vs. W	-0.38	0.13	0.04	-0.09	0.26	0.31	-0.83	-0.80	-0.75	0.02	-0.02	0.56	-0.04	-0.02	0.15
B vs. H/L	-0.65	-0.38	-0.09	-0.31	0.03		-0.83	-0.82	-0.83	0.52	0.03	0.56	-0.08	0.03	0.08
B vs. W	-0.48	-0.16	0.16	-0.18	0.30	0.30	-0.79	-0.72	-0.69	0.04	0.24	0.68	0.47	0.24	0.25
H/L vs. W	-0.40	-0.13	-0.01	-0.07	0.19		-0.91	-0.89	-0.87	-0.17	-0.28	0.34	-0.29	-0.28	-0.16

Table 8: Kendall's τ correlation coefficients for top items of different features. Missing entries are due to a lack of unique emoticons for the Hispanic/Latinx group.

annotations to binary 'yes' and 'no' by taking majority voting and discarding 7 users who were majority 'unsure,' the best self-report score model achieves 72.4% accuracy on the 192 users.

C MODEL HYPERPARAMETER AND TRAINING DETAILS

Our name model uses a CNN implementation released in Wood-Doughty et al. [68]. We use a CNN with 256 filters of width 3. The user's name (not screen name) is truncated at 50 characters and embedded into a 256 dimensional character embedding. We fine-tuned the learning rate on our dev data, trained for 250 epochs, and used early-stopping on dev-set F1 to pick which model to evaluate on the test set.

Our unigram model follows Volkova and Bachrach [66], using a simple sparse logistic regression. We use an implementation from Scikit-Learn, and tune the regularization parameter on the dev set. We introduce a hyperparameter to down-weight the contribution of our users compared to the baseline users; we also set that parameter on the dev set.

D REBALANCING ALGORITHM

All Twitter demographics datasets considered in this and previous work, unless they were explicitly class-balanced, have major class imbalances. This creates challenges when training on a dataset with one label distribution and then evaluating on a dataset with a different distribution. Variations of this challenge have been widely studied [41, 48]. To address this, after training our classifier we explicitly calculate its per-class recall on the development set. Then, we order each class from most to least-frequent, and adjust a threshold for predicting that class that should maximize total overall accuracy given the known label distribution of the dev or test set. Explicit details are in our released code.

Table 9 below shows that without our rebalancing approach, the accuracy of all trained models decreases, irrespective of the data on which they were trained. However, the F1 score of some models increases compared to the trained models which were rebalanced to target an imbalanced test set, because the model will make more predictions for low-probability labels.

	Names		Unigrams	
Dataset/Baseline	F1	Acc %	F1	Acc %
Random baseline	.250	25.0	.250	25.0
Majority baseline	.224	**80.8**	.224	80.8
Crowdsourced	.272	74.4	.396	**83.3**
GP	**.342**	72.6	.350	75.6
Crowdsourced+GP	.313	69.5	.426	81.8
WG	.335	64.3	.345	75.6
Crowdsourced+WG	.198	54.0	.406	79.6
BGP	.299	48.2	.304	43.8
Crowdsourced+BGP	.250	35.8	**.456**	76.8

Table 9: Imbalanced task prediction, without rebalancing. Accuracy decreases for all trained models, but F1 increases in some cases.

Using this method at test time requires prior knowledge on the *cumulative* label distribution which in many cases may be known due to high-level survey data [67], even when nothing is known about individuals in an dataset of interest. When we have no information on the true label distribution, this approach could be replaced with a more flexible Bayesian method, or we could abandon the rebalancing algorithm to avoid introducing any bias.

E DATA COLLECTION FOR GENDER SELF-REPORTS

It is notable that our data collection method could be extended to other demographics traits like gender and age, and it is also flexible with the label selection. We did a initial trial on gender data collection following the framework described above, with query regex:

```
\b((wo)?man|uncle|male|guy|dude|dad|father|brother|
husband|boy|papa|mr\.?]|mister|girl|lady|mom|mama|
wife|mother|sister|gal|girl|mr?s\.?)\b
```

Besides the plural and quote filter, we also explored two ad-hoc techniques. First, we filtered out query affected by possessive pronouns. Second, ignore the query in a frequently used self report pattern that causes many false positive in the collection like "daughter" in "dad of a lovely daughter". To evaluate the gender dataset, three annotators independently labeled 200 users that matched the gender queries, with Krippendorff's alpha of 0.8130. The matching for gender less noisy than the one for race. 68% users self report their gender as the matched query. We were able to collect a dataset of 207k users, with 90% precision on the manually labeled dataset.

REFERENCES

[1] Silvio Amir, Mark Dredze, and John W. Ayers. 2019. Population Level Mental Health Surveillance over Social Media with Digital Cohorts. In *NAACL Workshop on Computational Linguistics and Clinical Psychology*.

[2] Julia Angwin and Terry Parris Jr. 2016. Facebook lets advertisers exclude users by race. *ProPublica blog* 28 (2016).

[3] John W Ayers, Benjamin M Althouse, and Mark Dredze. 2014. Could behavioral medicine lead the web data revolution? *Jama* 311, 14 (2014), 1399–1400.

[4] Charley Beller, Rebecca Knowles, Craig Harman, Shane Bergsma, Margaret Mitchell, and Benjamin Van Durme. 2014. I'm a belieber: Social roles via self-identification and conceptual attributes. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vol. 2. 181–186.

[5] Katie Benner, Glenn Thrush, and Mike Isaac. 2019. Facebook Engages in Housing Discrimination With Its Ad Practices, US Says. *The New York Times* 28 (2019), 2019.

[6] Adrian Benton, Glen Coppersmith, and Mark Dredze. 2017. Ethical Research Protocols for Social Media Health Research. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. Association for Computational Linguistics, Valencia, Spain, 94–102. https://doi.org/10.18653/v1/W17-1612

[7] Brent Berlin and Paul Kay. 1991. *Basic color terms: Their universality and evolution.* Univ of California Press.

[8] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".

[9] Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic dialectal variation in social media: A case study of African-American English. *arXiv preprint arXiv:1608.08868* (2016).

[10] John D Burger, John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating gender on Twitter. In *EMNLP*. Association for Computational Linguistics, 1301–1309.

[11] Pete Burnap and Matthew L Williams. 2016. Us and them: identifying cyber hate on Twitter across multiple protected characteristics. *EPJ Data Science* 5, 1 (2016), 11.

[12] Mary E Campbell and Christabel L Rogalin. 2006. Categorical imperatives: The interaction of Latino and racial identification. *Social Science Quarterly* 87, 5 (2006), 1030–1052.

[13] Jonathan Chang, Itamar Rosenn, Lars Backstrom, and Cameron Marlow. 2010. epluribus: Ethnicity on social networks. In *Fourth International AAAI Conference on Weblogs and Social Media.*

[14] Kenneth Ward Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics* 16, 1 (1990), 22–29.

[15] R Dawn Comstock, Edward M Castillo, and Suzanne P Lindsay. 2004. Four-year review of the use of race and ethnicity in epidemiologic and public health research. *American journal of epidemiology* 159, 6 (2004), 611–619.

[16] Glen Coppersmith, Mark Dredze, and Craig Harman. 2014. Quantifying mental health signals in Twitter. In *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality.* 51–60.

[17] Stephen Cornell and Douglas Hartmann. 2006. *Ethnicity and race: Making identities in a changing world.* Sage Publications.

[18] Lorraine Culley. 2006. Transcending transculturalism? Race, ethnicity and health-care. *Nursing Inquiry* 13, 2 (2006), 144–153.

[19] Aron Culotta, Nirmal Ravi Kumar, and Jennifer Cutler. 2015. Predicting the Demographics of Twitter Users from Website Traffic Data.. In *AAAI.* 72–78.

[20] Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. *arXiv preprint arXiv:1306.6078* (2013).

[21] Laure Delisle, Alfredo Kalaitzis, Krzysztof Majewski, Archy de Berker, Milena Marin, and Julien Cornebise. 2019. A large-scale crowdsourced analysis of abuse against women journalists and politicians on Twitter. *CoRR* abs/1902.03093 (2019). arXiv:1902.03093 http://arxiv.org/abs/1902.03093

[22] Laure Delisle, Alfredo Kalaitzis, Krzysztof Majewski, Archy de Berker, Milena Marin, and Julien Cornebise. 2019. A large-scale crowdsourced analysis of abuse against women journalists and politicians on Twitter. *arXiv preprint arXiv:1902.03093* (2019).

[23] Virgile Landeiro Dos Reis and Aron Culotta. 2015. Using matched samples to estimate the effects of exercise on mental health via twitter. In *Twenty-Ninth AAAI Conference on Artificial Intelligence.*

[24] Mark Dredze, Michael J Paul, Shane Bergsma, and Hieu Tran. 2013. Carmen: A Twitter Geolocation System with Applications to Public Health. In *AAAI Workshop on Expanding the Boundaries of Health Informatics Using AI (HIAI)*.

[25] William W Dressler, Kathryn S Oths, and Clarence C Gravlee. 2005. Race and ethnicity in public health research: models to explain health disparities. *Annu. Rev. Anthropol.* 34 (2005), 231–252.

[26] Jacob Eisenstein, Amr Ahmed, and Eric P Xing. 2011. Sparse additive generative models of text. (2011).

[27] Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric P Xing. 2014. Diffusion of lexical change in social media. *PloS one* 9, 11 (2014), e113114.

[28] Casey Fiesler and Nicholas Proferes. 2018. 'Participant' Perceptions of Twitter Research Ethics. *Social Media+ Society* 4, 1 (2018), 2056305118763366.

[29] Lucie Flekova, Jordan Carpenter, Salvatore Giorgi, Lyle Ungar, and Daniel Preoţiuc-Pietro. 2016. Analyzing biases in human perception of user age and gender from text. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1. 843–854.

[30] A Gonzalez-Barrera and MH Lopez. 2015. Is being Hispanic a matter of race, ethnicity or both? *Pew Research Center* (2015).

[31] Charles Hirschman, Richard Alba, and Reynolds Farley. 2000. The meaning and measurement of race in the US census: Glimpses into the future. *Demography* 37, 3 (2000), 381–393.

[32] Arnold K Ho, Steven O Roberts, and Susan A Gelman. 2015. Essentialism and racial bias jointly contribute to the categorization of multiracial individuals. *Psychological Science* 26, 10 (2015), 1639–1645.

[33] Xiaolei Huang and Michael J Paul. 2018. Examining temporality in document classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 694–699.

[34] Isa Inuwa-Dutse, Bello Shehu Bello, and Ioannis Korkontzelos. 2018. Lexical analysis of automated accounts on Twitter. *arXiv preprint arXiv:1812.07947* (2018).

[35] Nicholas A Jones and Amy Symens Smith. 2001. *The two or more races population, 2000.* Vol. 8. US Department of Commerce, Economics and Statistics Administration, US.

[36] Rebecca Knowles, Josh Carroll, and Mark Dredze. 2016. Demographer: Extremely Simple Name Demographics. *NLP+ CSS 2016* (2016), 108.

[37] Gloria L Krahn, Deborah Klein Walker, and Rosaly Correa-De-Araujo. 2015. Persons with disabilities as an unrecognized health disparity population. *American journal of public health* 105, S2 (2015), S198–S206.

[38] Virgile Landeiro and Aron Culotta. 2016. Robust text classification in the presence of confounding bias. In *Thirtieth AAAI Conference on Artificial Intelligence*.

[39] Thomas A LaVeist. 2005. *Minority populations and health: An introduction to health disparities in the United States.* Vol. 4. John Wiley & Sons.

[40] Sharon M Lee and Sonya M Tafoya. 2006. Rethinking US census racial and ethnic categories for the 21st century. *Journal of Economic and Social Measurement* 31, 3-4 (2006), 233–252.

[41] Guillaume Lemaître, Fernando Nogueira, and Christos K Aridas. 2017. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *The Journal of Machine Learning Research* 18, 1 (2017), 559–563.

[42] Kate Loveys, Jonathan Torrez, Alex Fine, Glen Moriarty, and Glen Coppersmith. 2018. Cross-cultural differences in language markers of depression online. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*. 78–87.

[43] Alice E Marwick and Danah Boyd. 2011. I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience. *New media & society* 13, 1 (2011), 114–133.

[44] Jean-Baptiste Michel, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew K Gray, Joseph P Pickett, Dale Hoiberg, Dan Clancy, Peter Norvig, Jon Orwant, et al. 2011. Quantitative analysis of culture using millions of digitized books. *science* 331, 6014 (2011), 176–182.

[45] Jude Mikal, Samantha Hurst, and Mike Conway. 2016. Ethical issues in using Twitter for population-level depression monitoring: a qualitative study. *BMC medical ethics* 17, 1 (2016), 22.

[46] Burt L Monroe, Michael P Colaresi, and Kevin M Quinn. 2008. Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict. *Political Analysis* 16, 4 (2008), 372–403.

[47] Fred Morstatter, Jürgen Pfeffer, Huan Liu, and Kathleen M Carley. 2013. Is the sample good enough? comparing data from twitter's streaming api with twitter's firehose. In *Seventh international AAAI conference on weblogs and social media*.

[48] Vladimir Nikulin and Geoffrey J McLachlan. 2009. Classification of imbalanced marketing data with balanced random sets. In *Proceedings of the 2009 International Conference on KDD-Cup 2009-Volume 7*. JMLR. org, 89–100.

[49] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R Routledge, and Noah A Smith. 2010. From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM* 11, 122-129 (2010), 1–2.

[50] Michael J Paul and Mark Dredze. 2011. You are what you Tweet: Analyzing Twitter for public health. *Icwsm* 20 (2011), 265–272.

[51] Michael J Paul and Mark Dredze. 2017. Social monitoring for public health. *Synthesis Lectures on Information Concepts, Retrieval, and Services* 9, 5 (2017), 1–183.

[52] Umashanthi Pavalanathan and Jacob Eisenstein. 2015. Confounds and consequences in geotagged Twitter data. *arXiv preprint arXiv:1506.02275* (2015).

[53] Ellie Pavlick and Joel Tetreault. 2016. An empirical analysis of formality in online communication. *Transactions of the Association for Computational Linguistics* 4 (2016), 61–74.

[54] Marco Pennacchiotti and Ana-Maria Popescu. 2011. A Machine Learning Approach to Twitter User Classification. *Icwsm* 11, 1 (2011), 281–288.

[55] Daniel Preoţiuc-Pietro, Sharath Chandra Guntuku, and Lyle Ungar. 2017. Controlling human perception of basic user traits. In *Proceedings of the 2017 conference on empirical methods in natural language processing*. 2335–2341.

[56] Daniel Preoţiuc-Pietro and Lyle Ungar. 2018. User-Level Race and Ethnicity Predictors from Twitter Text. In *Proceedings of the 27th International Conference on Computational Linguistics*. 1534–1545.

[57] Daniel Preoţiuc-Pietro, Svitlana Volkova, Vasileios Lampos, Yoram Bachrach, and Nikolaos Aletras. 2015. Studying user income through language, behaviour and affect in social media. *PloS one* 10, 9 (2015), e0138717.

[58] Afshin Rahimi, Trevor Cohn, and Timothy Baldwin. 2016. pigeo: A Python Geotagging Tool. In *Proceedings of ACL-2016 System Demonstrations*. Association for Computational Linguistics, Berlin, Germany, 127–132. https://doi.org/10.18653/v1/P16-4022

[59] Juan Ramos et al. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, Vol. 242. Piscataway, NJ, 133–142.

[60] Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta. 2010. Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*. ACM, 37–44.

[61] David Satcher. 2001. Mental health: Culture, race, and ethnicityÎíñupplement to mental health: A report of the surgeon general.

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

[63] Donald P Spence and Kimberly C Owens. 1990. Lexical co-occurrence and association strength. *Journal of Psycholinguistic Research* 19, 5 (1990), 317–330.

[64] Fiona J Tweedie and R Harald Baayen. 1998. How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities* 32, 5 (1998), 323–352.

[65] Nicholas Vargas and Kevin Stainback. 2016. Documenting contested racial identities among self-identified Latina/os, Asians, Blacks, and Whites. *American Behavioral Scientist* 60, 4 (2016), 442–464.

[66] Svitlana Volkova and Yoram Bachrach. 2015. On predicting sociodemographic traits and emotions from communications in social networks and their implications to online self-disclosure. *Cyberpsychology, Behavior, and Social Networking* 18, 12 (2015), 726–736.

[67] Stefan Wojcik and Adam Hughes. 2019. Sizing Up Twitter Users. *Washington, DC: Pew Internet & American Life Project. Retrieved May* 1 (2019), 2019.

[68] Zach Wood-Doughty, Praateek Mahajan, and Mark Dredze. 2018. Johns Hopkins or johnny-hopkins: Classifying Individuals versus Organizations on Twitter. In *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*. Association for Computational Linguistics, New Orleans, Louisiana, USA, 56–61. https://doi.org/10.18653/v1/W18-1108

[69] Zach Wood-Doughty, Michael Smith, David Broniatowski, and Mark Dredze. 2017. How Does Twitter User Behavior Vary Across Demographic Groups?. In *Proceedings of the Second Workshop on NLP and Computational Social Science*. 83–89.