User-Level Race and Ethnicity Predictors from Twitter Text

Daniel Preoţiuc-Pietro
Computer and Information Science
University of Pennsylvania
danielpr@sas.upenn.edu

Lyle Ungar
Computer and Information Science
University of Pennsylvania
ungar@cis.upenn.edu

Abstract

User demographic inference from social media text has the potential to improve a range of downstream applications, including real-time passive polling or quantifying demographic bias. This study focuses on developing models for user-level race and ethnicity prediction. We introduce a data set of users who self-report their race/ethnicity through a survey, in contrast to previous approaches that use distantly supervised data or perceived labels. We develop predictive models from text which accurately predict the membership of a user to the four largest racial and ethnic groups with up to .884 AUC and make these available to the research community.

1 Introduction

The popularity and ubiquity of social media allows access to a wide variety of spontaneous language enabling researchers to study language variation across space and time at large scale (Eisenstein et al., 2010; Eisenstein, 2016). Through language analysis, social media data showed the potential to compliment or in part replace traditional polling (O'Connor et al., 2010), with the caveat that its demographics is not a representative sample of the real population (Culotta, 2014). One of the most important demographic differences – especially across the US – is that of race and ethnicity. For example, in Twitter-based political polling applications it is important to adjust for the fact that ethnic minorities are overall less likely to support either party, but prefer the Democratic Party over the Republican Party (Hajnal and Lee, 2011).

In this paper, we present the first extensive study on identifying user-level race and ethnicity. So far, linguistic differences have mostly been studied in the context of dialects, usually African-American Vernacular English – AAVE (Jørgensen et al., 2015), using message-level data which offered insight into syntactic (Stewart, 2014) or lexical markers (Blodgett et al., 2016). However, not all users from a racial or ethnic group use these markers or, more generally, an associated dialect and usage is different across socio-demographic traits – use of the AAVE is correlated with lower income and education (Rickford, 1999). In addition, there are differences in language use across racial/ethnic groups not caused by dialects e.g., in the US, African Americans prefer basketball, Whites ice-hockey and Hispanic/Latinos football (OpenDorse, 2013) – and are consequently more likely to post about these sports.

In addition, previous studies used either perceived race labels (Mohammady and Culotta, 2015; Volkova and Bachrach, 2016; Culotta et al., 2016) which are subject to human stereotypes (Flekova et al., 2016a) or mapped geo-located tweets to census statistics (Mohammady and Culotta, 2014; Blodgett et al., 2016) which lead to other biases (Jørgensen et al., 2015) because: 1) census statistics are outdated; 2) the Twitter population is not a representative sample of the general population (Eisenstein et al., 2011) with African Americans over-represented on Twitter (Duggan, 2015); 3) users who geo-locate tweets are not a representative sample of the Twitter population (Eisenstein, 2016); 4) geo-located tweets might be posted from a different location than the user's home.

In this study, we introduce a new data set where user-level race and ethnicity labels are collected through an online survey of Twitter users. We build models for accurately classifying Twitter users into four of the largest racial and ethnic groups in the U.S. as defined by the Census Bureau: Non-Hispanic Whites,

This work is licensed under a Creative Commons Attribution 4.0 International License. License details: http://creativecommons.org/licenses/by/4.0/.

Proceedings of the 27th International Conference on Computational Linguistics, pages 1534–1545
Santa Fe, New Mexico, USA, August 20-26, 2018.

Hispanic/Latinos, African-Americans and Asians.[1] We also measure the predictive power of various types of linguistic features and quantify usage of race/ethnicity revealing language across demographic groups. Finally, we perform a linguistic feature analysis to reveal the most important linguistic markers of race/ethnicity.

2 Data Set

We build a data set of Twitter users from participants in larger surveys taken through Qualtrics,[2] the largest platform for online experiments in social science, for which each user was compensated with 3$ per study.

As a first step, all participants were asked to take a standard demographic questionnaire which included selecting their race/ethnicity with the following options consistent with the US Census: African-American (AA, 374 users), Hispanic/Latino (Latino, 241 users), Asian (140 users), Non-Hispanic White (White, 3,182 users), Multiracial (153 users) or Other (free-text field prompt; 40 users, most frequently mentioned to be 'Native American'). In addition, we collected gender ('Male', 'Female' or open-ended field), age (13–90 interval), education (6 ordinal values ranging from 'No high-school' to 'Advanced Degree') and income level (8 ordinal values representing annual income from '<20k$/y' to '>200k$/y'). We restricted participation in our surveys to users based in the U.S. to limit the impact of potential cultural factors by using Qualtrics' filtering mechanisms.

The participants were asked to provide their public Twitter handle from which they had posted more than 100 tweets. We performed several checks on this Twitter handle to ensure that it was the user's own: *after* compensation we asked the users if they were truthful in reporting their handle and if not, we removed their data from analysis (22 users). We manually checked all handles which were verified by Twitter or had over 5000 followers and eliminated them if they were celebrities, as these were unlikely the users who participated in the survey (10 users in total). For each user, we downloaded their most recent 3,200 tweets, leading to a total data set size of 5,415,985 tweets.

We asked users to disclose their information, as this is the most efficient way to collect multiple demographic traits of users, despite the small possibility that some users are deceitful in reporting these traits. For ethical reasons, we only recruited participants who were willing to share their Twitter handles. This means our sample may not be fully representative of the general population, although we will account for any possible demographic skew in all our experiments.

After reporting demographics (including race/ethnicity) and their Twitter handles, users were directed to one of four collections of psychological questionnaires. The list of questions for each of these collections is very long (>30 questions) and bears no impact on the initial demographic information provided or to our experiments, hence we omit them from this paper. We have more 'White' users than the general U.S. population because participants were required to be part of this group in one of the sub-studies.

For the rest of the paper, we remove from our analysis the 'Multiracial' (153 users) and 'Other' (40 users) groups as these are small in sample size, heterogeneous in make-up and the sample may have been skewed towards a certain race/ethnicity mixture, but we could not know which.

3 About Race/Ethnicity

In our study, we used the definition of race/ethnicity as presented in the U.S. Census.[3,4]

We have chosen the U.S. Census definition of race/ethnicity for the following reasons:

- it is arguably most familiar to the participants, as users are all from the U.S., thus most have participated in the Census previously;
- multiple previous studies and applications such as polls rely on this classification to study the relationship between race/ethnicity and other variables;
- previous work on race-specific language used the same racial categorization (Blodgett et al., 2016);
- U.S. Census race/ethnicity surname statistics are available to benchmark our methods.

[1] Please refer to the 'About Race/Ethnicity' for a discussion about the definition of race/ethnicity used in this study.
[2] https://www.qualtrics.com/
[3] https://www.census.gov/topics/population/race/about.html
[4] https://www.census.gov/topics/population/hispanic-origin/about.html

We acknowledge that other racial/ethnic classifications exist or some groups are heterogeneous (e.g., Asian contains Chinese, Indian and Middle Eastern ethnicities) and that some argue the validity of the race construct entirely (Mukhopadhyay et al., 2013; Sen and Wasow, 2016).

4 Ethical Considerations

This experiment has received ethics approval from the Institutional Review Board (IRB) of our institution.

For reproducibility purposes, the data used in this paper is released in an anonymized and aggregated format.[5]. Any academic who wishes to gain access to the original user id's and labels will need to register with the authors, as per our IRB terms. Access will not be granted for any commercial applications and purposes.

Any individual-level predictions such as gender, age, religion, race or politics should be used with caution as they represent sensitive information. Race is, however, perhaps even more fraught with potential misunderstandings and abuses. We expect models for race prediction to be useful at an aggregate level, as the performance of the models do not warrant being be used at an individual level. For example, one could study how, in aggregate, tweets from youth of different races tweet about drug and alcohol use, where age and race make-up is estimated using predictive models.

We also highlight that Twitter's Terms of Service disallow targeting of individual users based on sensitive traits, including race: 'our Twitter Ads Policy also prohibits advertisers from targeting ads based on categories we consider sensitive, such as race, religion, politics, sex life, or health'.[6,7]

5 Features

In our analysis and predictive experiments, we use a broad range of linguistic features which are briefly described below.

Unigrams We use the bag-of-words representation to reduce each user's posting history to a normalised frequency distribution over the vocabulary consisting of all words used by at least 1% of the users (32,142 words).

LIWC Traditional psychological studies use a dictionary-based approach to representing text. The most popular method is based on Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2001) consisting of 73 manually constructed lists of words (Pennebaker et al., 2015) including some specific parts-of-speech, topical or stylistic categories. Each user is thereby represented as a frequency distribution over these categories.

Word2Vec Topics An alternative to LIWC is to use automatically generated word clusters. These clusters of words can be thought of as *topics*, i.e., groups of words that are semantically and/or syntactically similar. The clusters help reduce the feature space and provide good interpretability. To create these groups of words, we use an automatic method that leverages word co-occurrence patterns in large corpora by making use of the distributional hypothesis: similar words tend to co-occur in similar contexts (Harris, 1954).

We use the method from (Preoţiuc-Pietro et al., 2015a) to compute topics using word2vec similarity (Mikolov et al., 2013) and spectral clustering (Shi and Malik, 2000; von Luxburg, 2007) of different sizes (from 30 to 2000). We have tried other alternatives to building clusters: using other word similarities to generate clusters – such as NPMI (Lampos et al., 2014) or GloVe (Pennington et al., 2014) as proposed in (Preoţiuc-Pietro et al., 2015a) – or using standard topic modelling approached to create soft clusters of words e.g., Latent Dirichlet Allocation (Blei et al., 2003). For brevity and clarity, we present experiments with the best performing feature set containing 1000 Word2Vec topics after experimenting with other numbers of topics (from 30 through 2000). We aggregated all the words posted in a users tweets and represent each user as a distribution of the fraction of words belonging to each cluster.

Sentiment & Emotions We also investigate the extent to which racial groups differ in the type of emotions they express through their posts. The most popular model of discrete emotions is the Ekman model (Ekman, 1992; Strapparava and Mihalcea, 2008; Strapparava et al., 2004) which posits the existence of six basic

[5] http://www.preotiuc.ro
[6] https://twitter.com/privacy/
[7] https://support.twitter.com/articles/20170368\#/

emotions: anger, disgust, fear, joy, sadness and surprise. We automatically quantify these emotions from our Twitter data set using a publicly available model, which in addition to emotions also includes positive and negative sentiment (Volkova and Bachrach, 2016). We also used the word lexicon derived using crowd-sourcing from (Mohammad and Turney, 2010; Mohammad and Turney, 2013) and found it had slightly worse predictive results. Using these models, we assign sentiment and emotion probabilities to each message and then average across all users' posts to obtain user level emotion expression scores.
Part-of-Speech Tags We analyze part-of-speech tag usage across groups by POS tagging all tweets using the Twitter model of the Stanford Tagger (Derczynski et al., 2013), which showed best tagging results for AAVE (Jørgensen et al., 2015) and also uses the finer grained Penn Treebank tagset. Each user is thus represented as a distribution over POS tags.
Linear Ensemble Finally, we build a logistic regression model having as features the real-valued predictions of the models trained on all previous feature sets.

6 Baselines

We introduce the following competitive baselines:
Demographics User demographics collected from the users using our survey (age, gender, education and income level) are used as features. A classifier using only these demographics as features will establish the demographic tendencies of race/ethnicity in the data set.
Blodgett – User Model Blodgett et al. (2016) release a set of tweets with race/ethnicity probabilities obtained through distant supervision. In order to create user-level models for comparison, we take two separate approaches. From that data set, we identify the users that have more than 50 messages in the data set and label them with one of the four groups (AA, Latino, Other – mostly Asian – and White) if more than half of their messages are assigned a probability higher than 0.8 for that group. We build four one-vs-all unigram user-level classifiers and use these four models to predict the race/ethnicity of the users in our data set. These models were internally validated on the data set from (Blodgett et al., 2016) obtaining an average AUC $= .970$ prediction accuracy on 10-fold cross-validation within that data set. We also note that this method is trained on data from 26,009 users, an order of magnitude larger than our data set.
Blodgett – Message Model The second approach based on the (Blodgett et al., 2016) data set is to first train message-level logistic regression models for predicting how likely each message belongs to one of the four race/ethnicity groups. On internal validation, the average Pearson correlation between these models' prediction and the message-level scores released is R=.534 as measured using 10 fold cross-validation. We use these models to assign four prediction scores to all messages in our data set and compute an user-level average across these four predictions. We note that in this method we use 50 million tweets for training the message-level race/ethnicity classifiers, again an order of magnitude larger than our data set.
Perceived Race Labels We experiment with two data sets consisting of users for which race was determined by external annotators who analyzed the Twitter profile and tweets of users. The first data set was introduced in (Mohammady and Culotta, 2014) and consists of 362 users (all remaining public accounts from the original 770 users) labelled with three race categories (AA – 70, Latino – 104, White – 188). The second data set was introduced in (Volkova and Bachrach, 2016) and consists of 3380 users (all remaining public accounts from the original 5000 users) labelled with four race categories (AA – 1216, Latino – 150, Asian – 316, White – 1698) . All models trained on this data use bag-of-word unigram features and have been pre-processed in the same way as all other data sets.
Census Surnames Previous research used aggregate statistics about name distributions to estimate the demographic traits of Twitter users based on their name field, especially focusing on gender (Knowles et al., 2016). The U.S. 2010 Census aggregates statistics on race distribution for all surnames common to at least 100 US citizens, covering 95.9% of the population.[8] In total, 65.9% (2,592) of the users in our data set have a valid surname in their Twitter name field and could be matched to the Census statistics. For these, we selected the race with the highest frequency in the Census data as our prediction and for the rest, we use the most frequent race, namely White, as the prediction (**Census Surname – Most Likely**). We

[8] http://www.census.gov/topics/population/genealogy/data/2010_surnames.html

	AA	Latino	Asian	White
Baseline	.500	.500	.500	.500
Demographics	.630	.638	.659	.618
Models using Surnames				
Census Surname – Most Likely	.536	.610	.630	.571
Census Surname – Distribution	.634	.710	.736	.678
NamePrism - Most Likely	.525	.648	.657	.572
NamePrism - Distribution	.681	*.740*	.765	.719
Models based on (Blodgett et al., 2016)				
Blodgett – User Model	.729	.579	.560	.645
Blodgett – Message Model	.802	.700	.614	.718
Models Trained on Perceived Race Labels				
(Mohammady and Culotta, 2014)	.798	.603	–	.700
(Volkova and Bachrach, 2016)	.859	.693	.736	.765
User-level models proposed in this paper				
Emotions	.621	.617	.578	.609
LIWC	.775	.651	.656	.689
POS	.721	.666	.619	.686
Topics	.840	.670	.731	.758
Unigrams	.866	.708	.768	.795
Linear Ensemble	*.870*	*.710*	*.781*	*.797*
Linear Ensemble & NamePrism	**.884**	**.781**	**.832**	**.825**

Table 1: User-level race one-vs-all classification results measured in ROC AUC.

	Original		Balanced	
	Acc	F1	Acc	F1
Random Guess	.808	.223	.250	.100
Demographics	.809	.283	.391	.384
Census Surname – Most Likely	.804	.372	.358	.303
Census Surname – Distribution	.806	.403	.411	.411
NamePrism - Most Likely	.798	.386	.388	.342
NamePrism - Distribution	.801	.418	.480	.463
Trained on (Mohammady and Culotta, 2014)	.780	.338	.367	.287
Trained on (Volkova and Bachrach, 2016)	.820	.401	.383	.318
Trained on this data set	*.837*	*.485*	*.555*	*.557*
Trained on this data set & NamePrism – Distribution	**.842**	**.537**	**.617**	**.614**

Table 2: User-level four way race classification results on both the original data set and on balanced classes obtained by over-sampling the less frequent classes. Bag-of-words unigram features are used in all text-based methods.

also use the four Census race frequencies for the user's matched surname as features (**Census Surname – Distribution**). For users that are not matched, we use the average distribution. This has the effect to adjust for, at least partially, the race differences between the general U.S. population and the Twitter users in our data set.

NamePrism In addition to only looking at surnames as extracted from the Census, we also use a publicly available API [9] that enables race and nationality prediction from word formation patterns in surnames. This calls models that have obtained state-of-the-art accuracy for this task (Ye et al., 2017; Ambekar et al., 2009). Similarly to the Census name, we use both the actual prediction (**NamePrism – Most Likely**) and a model trained on the race probabilities predicted by the API (**NamePrism – Distribution**).

7 Predictive Experiments

In our predictive experiments we use logistic regression classification with Elastic Net regularization in a 10-fold cross-validation setup, where 8 folds are used for training, 1 for tuning the regularization parameters using grid search and 1 for testing. We have experimented with other methods for non-linear classification (e.g. SVMs), but results did not improve significantly.

7.1 One vs. All Classification

Results of one vs. all classifications are presented in Table 1. These are evaluated using ROC AUC (area under the receiver operating characteristic curve) because: a) the class distribution is imbalanced (up to

[9] http://www.name-prism.com/

1:19 for Asian vs. rest); b) AUC does not favor methods that use in training a similar class distribution as in the test set. Results using F1 score show the same patterns of improvement.

Overall performance of our best models are >.7 AUC (compared to AUC $= .5$ if random), with performance on the 'African American' group being highest (AUC $= .884$). Unigrams consistently obtain the best predictive results out of all feature sets, as they can capture idiosyncratic words and spellings which are very specific to certain racial/ethnic groups, while topic choice can only capture more generic topical relationships. Textual features – with a few exceptions for the 'Asian' group – make better predictions when compared to demographics, showing that language use carries information beyond purely demographics. Combining the predictions of all feature sets using a linear ensemble obtains the best results.

The two methods based on the data from (Blodgett et al., 2016) obtain overall good prediction results, with the message-level model surpassing the user-level model in all four cases. However, the performance is consistently lower than the best classifier trained on our data set, despite this being an order of magnitude smaller. Similar classifiers trained on perceived age labels result in models that are slightly, but consistently worse (.022 on average when comparing the 'Unigrams' only models) than when training on our data set. The results indicate that using perceived traits results in less accurate models when predicting real traits (Flekova et al., 2016a), as data set size is similar when using the data set from (Volkova and Bachrach, 2016). In addition, we have tried standard domain adaptation methods for combining both data sets (Daumé III, 2007) but found no additional gains.

The methods based on surnames obtain performance above chance, but below our best predictive models, except for the 'Latino' group where it exceeds our best model's performance due to the peculiarities posed by Hispanic/Latino surnames. Combining surname-based methods with the best text-based model results leads to significantly better results than using text alone, showing that these contain complementary information. However, further adding the demographics (age, gender, income, education) as features does not add to the predictive performance, showing that the results are not impacted by any imbalance in the demographic makeup of our data set.

7.2 Four-way Classification

Next, we experiment with four-way race/ethnicity classifiers using only unigram features for text-based methods, as these performed best in the previous experiment. In addition to experiments on the original data set, we also performed experiments with a balanced data set obtained by oversampling the smaller classes. In the 'Trained on this data set' setup, we oversampled in each fold and data split, such that no users from training are present in testing.

The results are presented in Table 2 with results showing both accuracy and F1 score (macro averaged). We notice a higher margin of improvement when using the data from our data set when compared perceived race labelled users. Surname distributions perform second best and are the only features that bring additional performance on top of the best performing method.

7.3 In-Sample Perceived Race Label Prediction

Finally, we compute the performance of models trained in-sample on perceived labels from previous work (Mohammady and Culotta, 2014; Volkova and Bachrach, 2016). All models are logistic regression classifiers with Elastic Net regularization. If these models achieve significantly higher accuracies compared to when tested on our data set, we can conclude that the race/ethnicity prediction task was over-simplified by the method of selecting users in the data set.

Results on in-sample 10-fold cross-validation on the two data sets with perceived race labels are presented for one-vs-all classification in Table 3. We notice that the results are very high compared to when applying the models on data with real race labels, with models suffering on average >.10 drop in AUC on data from (Volkova and Bachrach, 2016) and up to a .30 drop when trained on data from (Mohammady and Culotta, 2014).

Similarly, Table 4 shows 4-way classification results for in-sample perceived race prediction. We again note the larger margins of improvement over the baseline for both classifiers when compared to the same classifiers applied on real race labelled data in our data set in Table 2.

	AA	Latino	Asian	White
Baseline	.500	.500	.500	.500
Tested using Survey Labels				
(Mohammady and Culotta, 2014)	.798	.603	–	.700
(Volkova and Bachrach, 2016)	.859	.693	.736	.765
Tested using Perceived Labels				
(Mohammady and Culotta, 2014)	.954	.912	–	.928
(Volkova and Bachrach, 2016)	.946	.780	.767	.892

Table 3: Predictive results of unigram models using perceived race labels. Results are in ROC AUC.

	Original		Balanced	
	Acc	F1	Acc	F1
Random Guess	.519	.227	.333	.250
Trained on (Mohammady and Culotta, 2014)	.771	.770	.777	.768
Random Guess	.502	.167	.250	.100
Trained on (Volkova and Bachrach, 2016)	.764	.557	.639	.637

Table 4: User-level four way race unigram-based classification results on predicting perceived race labels. Results are on both the original data set split and on balanced classes obtained by over-sampling the less frequent classes.

8 Demographic Covariates

Next, we explore the hypothesis that the accuracy of race/ethnicity classification varies with the demographic traits of the users. For example, previous hypotheses state that AAVE usage is more prevalent in males, people of lower income and lower education levels (Rickford, 1999), thus making race prediction for this groups more accurate.

We focus our analysis on the African American user group. We build data splits that are matched in size, one demographic trait at a time (gender, age, education, income), by sub-sampling the larger demographic group when necessary. The, we sub-sample users such that the label proportions (AA vs. other) are the same in each task. For example, in studying the impact of gender in race classification we built a AA classifier for male users (122 AA vs. 122 Non-AA) and another one for female users (122 AA vs. 122 Non-AA). The split points between the two compared groups for age, education level and income level are 26 years old, completed High School degree and 40,000$/y respectively. All classifiers were trained using logistic regression with unigram features.

This experiment allows to uncover for which demographics automated methods are able to better classify users in belonging to the African American group, while keeping the training data size and label distribution constant. Classification results are shown in Table 5.

Classifiers reach different accuracies for age, gender and income, with younger, female and lower income users all being significantly easier to predict if they belong to the African American group. Further, in a separate experiment, the final classifier is more accurate for the same user categories (e.g., AUC $= .894$ for females compared to AUC $= .810$ for males). These results have implications in biases and classifier fairness.

Gender	Female: 0.858	Male: 0.772
Age	<26: 0.878	>26: 0.759
Education	≤High School: 0.819	>High School: 0.802
Income	<40,000$/y: 0.853	>40,000$/y: 0.770

Table 5: Classification performance in ROC AUC when holding data set size constant across different demographic traits.

9 Feature Analysis

Finally, we perform a linguistic feature analysis with the goal of identifying the linguistic markers specific of each race/ethnicity group on Twitter.

First, Table 6 shows the features with the highest weights for each race/ethnicity category. The features represent the weights learnt a regularized bag-of-words unigram logistic regression models trained for predicting if a user is part of a race/ethnicity group, as described in Section 7.1. The results are intuitive, highlighting known dialectal variations (e.g., 'bout', 'naw') and references to in-group figures ('trayvon',

Latino	Asian
latinas, barely, maze, latina, remorse, dd, absorbed, ay	asian, neighbourhood, pho, ud, jg, consciousness
White	**African-American**
of, seriously, unhappy, sure, pumped, great, someday	trayvon, meek, bout, yaaasss, beyonce, smh, naw

Table 6: Unigrams most predictive of each race/ethnicity group when examining the feature weights learned by the classifiers.

r	Cluster	Words	r	Cluster	Words
Latino			**Asian**		
.144 → .063		22 Significant Clusters with mostly Spanish words	.145→ .060		13 Significant Clusters with words in Asian languages
.148	Spanish City names	los, san, diego, angeles, francisco, antonio, jose, fran	.107	Asian Place-names	china, japan, korea, tokyo, asia, philippines, seoul, hk
.125	S.American Places	mexico, venezuela, brazil, puerto, rico, buenos, costa, argentina	.105	Elongated 'yes'	yah*, yaa*, yee*, wahh, hee
.088	Politeness	no, reason, matter, problem, sense, offense, worries, excuses	.098	Phonetic Spelling	ikr, rly, h8, bcuz, rlly, coz, pple, realy, ure, completely
.078	Frustration	dayum, man*, damn*, shi*t, ma*n	.087	Intejections	eh, ye, heh, tis, pun, poke, wink, teh, mam
.070	Love	333*, love*, lo*ve, ilove	.084	Cricket	england, india, kenya, #worldcup, cricket, batting, aus, pakistan
White			**African-American**		
.121	Time References	years, hours, minutes,ago, weeks, months, minute, seconds, min	.281	Slang	bout, wit, kno, sayin, talkin, doe, lowkey, wut, naw, thinkin
.110	Superlatives	ever, worst, biggest, cutest, funniest, coolest, longest, sweetest	.252	Group References	bitches, hoes, cops, dudes, fools, fucks, females, chicks, shits
.101	Adverbs of degree	absolutely, quite, extremely, perfectly, incredibly, certainly	.241	Person References	lil, boo, buddy, mama, sis, bby, tina, dee, missy, mister, sissy
.100	Rest	bed, couch, laying, cuddle, blanket, comfy, cuddling, blankets	.224	Phonetic Spelling	somethin, urself, some1, every1, any1, wha, sum1, no1
.091	Modals	say, could, wish, i'd, wouldn't, couldn't, meant, you'd, couldnt	.171	G-Dropping	tl, freaky, callin, actin, seein, tweetin, followin, ppls, sendin

Table 7: Word2Vec clusters most correlated with each race/ethnicity group (maximum 5 clusters per group, excluding topics made up of foreign words). The cluster name is manually assigned. All correlations are significant at $p < .05$, two-tailed t-test, **Simes corrected** for multiple comparisons and are controlled for gender, age, education and income. Words in a category are sorted by frequency in our data set. * highlights word variants where a character is repeated.

'beyonce') for AA users, names specific to a racial or ethnic group (e.g. 'latinas', 'latina', 'asian'), frequent foreign words (e.g. 'ay', 'ud') or more salient linguistic cues ('seriously', 'of', 'great', 'barely').

However, examining only the most predictive features assigned by a model with both L1 and L2 regularization is likely to overlook features are co-linear. For this reason, we perform analysis of Word2Vec clusters, POS Tags and Emotions using univariate correlation as introduced in (Schwartz et al., 2013). We compute univariate Pearson correlations independently for each feature between its distribution across users (features are first normalized to sum up to unit for each user) and the user-level outcome of interest (e.g., whether it belongs to a race/ethnicity group or not). In Section 8, we showed that user demographics (age, gender, education and income levels) impact the classification accuracies. In order to mitigate this effect, we introduce age, gender, education and income levels as controls and compute partial linear correlation for each feature. Since a large number of features are explored simultaneously, we consider coefficients significant if they meet a Simes-corrected two-tailed p-value of less than 0.05.

Table 7 shows the Word2Vec clusters with the highest correlations with users belonging to a group. For the 'Latino' and 'Asian' groups, we do not show clusters that contain more than 70% non-English words in their most frequent 20 words, as these only reflect the use of a particular language (e.g. Hindi or Tagalog for 'Asian' and Spanish for 'Latino') and are not interesting for analysis.

For Latinos we first note the use of Spanish-origin place names in the US and place names from Latin America. In addition, Latinos use words associated with polite constructs (e.g. 'offense' as in 'no offense', 'worries' as in 'no worries', 'excuses') and express frustration (variants of 'man', 'shit' and 'damn', as well as the Spanish version 'dayum') or love (heart shapes or variants of 'love'). The Asian group is, similarly to the 'Latino' group, first characterized by the use of words belonging to Asian languages and Asian name places. Asian users prefer a peculiar type of expressing 'yes' (e.g.,' 'yah', 'yaa', 'yee'), use a specific category of interjections (e.g., 'heh', 'pole') and prefer certain phonetic spellings (e.g., 'bcuz', 'h8') or contractions to typical English words (e.g. ikr – 'I know, right?', 'rly', 'rlly' – 'really'). Finally,

r	Tag	Words	r	Tag	Words
Latino			**Asian**		
.100	FW	Foreign Word	.079	UH	Interjection
.097	UH	Interjection	.057	NNP	Proper Noun, singular
.091	NNP	Proper Noun, singular			
White			**African-American**		
.101	IN	Subordinating conj. – of, on	.098	RP	Adverb, particle – about
.093	JJS	Adjective, superlative – best	.097	VB	Verb, base form – think
.087	DT	Determiner – the	.094	PRP	Personal Pronoun – I
.085	EX	Existential 'there'	.084	WP	Wh-pronoun – Who
.083	VBZ	Verb, 3rd pers. singular present – moves	.053	VBP	Verb, non-3rd pers. singular present – move
.082	TO	to	.051	JJR	Adjective, comparative – big
.069	MD	Modal – could			
.058	RB	Adverb – extremely			
.054	WDT	Wh-determiner – which			
.054	CC	Coordinating conj. – and			
.044	NNS	Noun, plural – years			

Table 8: Part-of-Speech tags most correlated with each race/ethnicity group. All correlations are significant at $p < .05$, two-tailed t-test, **Simes corrected** for multiple comparisons and are controlled for gender, age, education and income.

Emotion	Latino	Asian	White	AA
Positive	-.060	–	–	.036
Negative	-.049	–	–	–
Anger	–	–	-.067	.118
Disgust	–	–	–	.083
Fear	–	–	–	.050
Joy	-.044	–	–	–
Sadness	-.040	–	–	.048
Surprise	-.041	–	–	–

Table 9: Pearson correlations between emotions and race/ethnicity groups. All reported correlations are significant at $p < .05$, two-tailed t-test, **Simes corrected** for multiple comparisons and are controlled for gender, age, education and income.

the cluster containing words about cricket highlights a more topical difference specific of Asian users, likely driven by users with ethnicity in the Indian subcontinent where this sport is popular. White users are best characterized by the use of temporal references (e.g., 'years', 'ago', 'weeks'). Many clusters uncover syntactic preferences of the White group including the use of superlatives (e.g., 'cutest'), adverbs of degree (e.g., 'quite') and modal verbs (e.g. 'could'). Finally, a topical cluster around the theme of rest (e.g., 'laying', 'cuddle') is also specific of the White group. Finally, the African American groups has the topics with the highest correlation coefficients. Clusters capture specific terms that reference other persons (e.g., 'lil', 'boo') or groups (e.g., 'dudes') and other typical AAVE words (e.g., 'naw', 'bout'). Similarly to the Asian group, we also observe a cluster dominated by phonetic spelling variants of words (e.g., 'urself', 'some1'). Finally, we observe a topic dominated by verbs missing the final 'g' (e.g., 'callin', 'sendin'). This represents one of the most well-known distinctive features of AAVE (Rickford, 1999).

Table 8 shows Part-of-Speech tags correlated with each race/ethnicity group. The Latino and Asian groups are characterized by the use of foreign words, interjections and singular proper nouns. This matches the findings of the cluster analysis, leading to the conclusion that these two groups use references of proper names and foreign words both related to their ethnic origin and, increased use of interjections. The POS usage results for the White and African-American groups highlight intriguing contrasts. While the White users prefer superlative adjectives, African American users prefer comparative adjectives. White users are predisposed to use more adverbs, however African Americans use more adverb particles (e.g., 'take off', 'put away'). African-American prefer non 3-rd person singular verbs and use more personal pronouns, while White users prefer 3-rd person verbs. This shows the tendency of African-Americans to post more on Twitter about their activities and whereabouts rather than other events or impersonal reporting. White users also use more conjunctions (subordinating and coordinating), perhaps an indicator of more complex syntactic constructions.

Lastly, emotion analysis results are presented in Table 9. The results show that emotions are mainly correlated with the Latino and AA groups, with only a single other significant correlation for the Asian and White groups. African American users on Twitter express overall more emotions than other groups, including both positive sentiment as well as a range of negative emotions, especially anger. On the other hand, Latino users express overall less sentiment, both positive and negative, and fewer emotions.

10 Conclusion

We presented a detailed study of user-level race/ethnicity prediction along the lines of previous work on predicting user traits from text (Burger et al., 2011; Rao et al., 2010; Pennacchiotti and Popescu, 2011; Schwartz et al., 2013; Sap et al., 2014; Volkova et al., 2014; Preoţiuc-Pietro et al., 2015b; Flekova et al., 2016b; Preoţiuc-Pietro et al., 2016; Preoţiuc-Pietro et al., 2017). In contrast with previous research on race/ethnicity, we used labels obtained by directly surveying Twitter users, rather than distantly supervised geo-located data or perceived race labels, which lead to multiple biases and lower accuracies on real data.

We built models that obtain state-of-the-art out-of-sample accuracy on predicting the four prominent racial/ethnic groups in the US. We presented an extensive linguistic feature analysis for each group and brought new evidence towards linguistic hypotheses on dialect use in Twitter across demographic groups. We believe this paper offers a solid basis for the study of race prediction online. Our models are readily usable for large-scale passive polling scenarios, where automatically quantifying existing racial sample differences can lead to improved predictive performance (Culotta, 2014).

Acknowledgments

The authors acknowledge the support of the Templeton Religion Trust, grant TRT-0048.

References

Anurag Ambekar, Charles Ward, Jahangir Mohammed, Swapna Male, and Steven Skiena. 2009. Name-ethnicity Classification from Open Sources. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge Discovery and Data Mining*, KDD, pages 49–58.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Su Lin Blodgett, Lisa Green, and Brendan O'Connor. 2016. Demographic Dialectal Variation in Social Media: A Case Study of African-American English. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, EMNLP, pages 1119–1130.

D. John Burger, John Henderson, George Kim, and Guido Zarrella. 2011. Discriminating Gender on Twitter. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, EMNLP, pages 1301–1309.

Aron Culotta, Nirmal Ravi Kumar, and Jennifer Cutler. 2016. Predicting Twitter User Demographics using Distant Supervision from Website Traffic Data. *Journal of Artificial Intelligence Research*, 55:389–408.

Aron Culotta. 2014. Reducing Sampling Bias in Social Media Data for County Health Inference. In *Joint Statistical Meetings Proceedings*, pages 1–12.

Hal Daumé III. 2007. Frustratingly Easy Domain Adaptation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, ACL, pages 256–263.

Leon Derczynski, Alan Ritter, Sam Clark, and Kalina Bontcheva. 2013. Twitter Part-of-Speech Tagging for All: Overcoming Sparse and Noisy Data. In *Proceedings of Recent Advances in Natural Language Processing*, RANLP, pages 198–206.

Maeve Duggan. 2015. *Mobile Messaging and Social Media – 2015*. Pew Research Center.

Jacob Eisenstein, Brendan O'Connor, Noah A. Smith, and Eric P. Xing. 2010. A Latent Variable Model for Geographic Lexical Variation. In *Proceedings of the 2010 Conference on Empirical Methods for Natural Language Processing*, EMNLP, pages 1277–1287.

Jacob Eisenstein, Noah A Smith, and Eric P Xing. 2011. Discovering Sociolinguistic Associations with Structured Sparsity. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics*, ACL, pages 1365–1374.

Jacob Eisenstein. 2016. Written Dialect Variation in Online Social Media. In Charles Boberg, John Nerbonne, and Dom Watt, editors, *Handbook of Dialectology*. Wiley.

Paul Ekman. 1992. An Argument for Basic Emotions. *Cognition & Emotion*, 6(3-4):169–200.

Lucie Flekova, Jordan Carpenter, Salvatore Giorgi, Lyle Ungar, and Daniel Preoţiuc-Pietro. 2016a. 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*, ACL, pages 843–854.

Lucie Flekova, Lyle Ungar, and Daniel Preoţiuc-Pietro. 2016b. Exploring Stylistic Variation with Age and Income on Twitter. In *Proceedings of the 54th annual meeting of the Association for Computational Linguistics*, ACL.

Zoltan L Hajnal and Taeku Lee. 2011. *Why Americans don't Join the Party: Race, Immigration, and the Failure (of Political Parties) to Engage the Electorate*. Princeton University Press.

Z. Harris. 1954. Distributional Structure. *Word*, 10(23):146 – 162.

Anna Katrine Jørgensen, Dirk Hovy, and Anders Søgaard. 2015. Challenges of Studying and Processing Dialects in Social Media. In *Proceedings of the First Workshop on Noisy User-Generated Text (W-NUT)*, ACL, pages 9–18.

Rebecca Knowles, Josh Carroll, and Mark Dredze. 2016. Demographer: Extremely Simple Name Demographics. In *Proceedings of the Workshop on NLP and Computational Social Science*, EMNLP, pages 108–113.

Vasileios Lampos, Nikolaos Aletras, Daniel Preoţiuc-Pietro, and Trevor Cohn. 2014. Predicting and Characterising User Impact on Twitter. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, EACL, pages 405–413.

Tomas Mikolov, Wen tau Yih, and Geoffrey Zweig. 2013. Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of the 2010 annual Conference of the North American Chapter of the Association for Computational Linguistics*, NAACL, pages 746–751.

Saif M. Mohammad and Peter D. Turney. 2010. Emotions Evoked by Common Words and Phrases: Using Mechanical Turk to Create an Emotion Lexicon. In *Proceedings of the Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, NAACL, pages 26–34.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a Word-Emotion Association Lexicon. *Computational Intelligence*, 29(3):436–465.

Ardehaly Ehsan Mohammady and Aron Culotta. 2014. Using County Demographics to Infer Attributes of Twitter Users. In *Proceedings of the Joint Workshop on Social Dynamics and Personal Attributes in Social Media*, ACL, pages 7–16.

Ardehaly Ehsan Mohammady and Aron Culotta. 2015. Inferring Latent Attributes of Twitter Users with Label Regularization. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL, pages 185–195.

Carol C Mukhopadhyay, Rosemary Henze, and Yolanda T Moses. 2013. *How Real is Race?: A Sourcebook on Race, Culture, and Biology*. Rowman & Littlefield.

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. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, ICWSM, pages 122–129.

OpenDorse. 2013. 2013 Sports Fan Demographics. `http://opendorse.com/blog/2013-sports-fan-demographics/`. Accessed: 2017-02-05.

Marco Pennacchiotti and Ana-Maria Popescu. 2011. A Machine Learning Approach to Twitter User Classification. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, ICWSM, pages 281–288.

James W. Pennebaker, Martha E. Francis, and Roger J. Booth. 2001. *Linguistic Inquiry and Word Count*. Mahway: Lawrence Erlbaum Associates.

James W. Pennebaker, Roger J. Booth, Ryan L. Boyd, and Martha E. Francis. 2015. *Linguistic Inquiry and Word Count: LIWC2015*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, EMNLP, pages 1532–1543.

Daniel Preoţiuc-Pietro, Vasileios Lampos, and Nikolaos Aletras. 2015a. An Analysis of the User Occupational Class through Twitter Content. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, ACL, pages 1754–1764.

Daniel Preoţiuc-Pietro, Svitlana Volkova, Vasileios Lampos, Yoram Bachrach, and Nikolaos Aletras. 2015b. Studying User Income through Language, Behaviour and Affect in Social Media. *PLoS ONE*, 10(9), 09.

Daniel Preoţiuc-Pietro, Jordan Carpenter, Salvatore Giorgi, and Lyle Ungar. 2016. Studying the Dark Triad of Personality using Twitter Behavior. In *Proceedings of the 25th ACM Conference on Information and Knowledge Management*, CIKM, pages 761–770.

Daniel Preoţiuc-Pietro, Ye Liu, Daniel J. Hopkins, and Lyle Ungar. 2017. Beyond Binary Labels: Political Ideology Prediction of Twitter Users. In *Proceedings of the 55th Conference of the Association for Computational Linguistics*, ACL.

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*, SMUC, pages 37–44.

John Russell Rickford. 1999. *African American Vernacular English: Features, Evolution, Educational Implications*. Wiley-Blackwell.

Maarten Sap, Gregory Park, Johannes Eichstaedt, Margaret Kern, Lyle Ungar, and H Andrew Schwartz. 2014. Developing Age and Gender Predictive Lexica over Social Media. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, EMNLP, pages 1146–1151.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, and Martin EP Seligman. 2013. Personality, Gender, and Age in the Language of Social Media: The Open-vocabulary Approach. *PloS ONE*, 8(9).

Maya Sen and Omar Wasow. 2016. Race as a Bundle of Sticks: Designs that Estimate Effects of Seemingly Immutable Characteristics. *Annual Review of Political Science*, 19:499–522.

Jianbo Shi and Jitendra Malik. 2000. Normalized Cuts and Image Segmentation. *Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905.

Ian Stewart. 2014. Now we Stronger than Ever: African-American Syntax in Twitter. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, EACL, pages 31–37.

Carlo Strapparava and Rada Mihalcea. 2008. Learning to Identify Emotions in Text. In *Proceedings of the 2008 ACM Symposium on Applied Computing*, pages 1556–1560.

Carlo Strapparava, Alessandro Valitutti, et al. 2004. WordNet Affect: an Affective Extension of WordNet. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, volume 4 of *LREC*, pages 1083–1086.

Svitlana Volkova and Yoram Bachrach. 2016. Inferring Perceived Demographics from User Emotional Tone and User-Environment Emotional Contrast. In *Proceedings of the 54rd Annual Meeting of the Association for Computational Linguistics*, ACL, pages 1567–1578.

Svitlana Volkova, Glen Coppersmith, and Benjamin Van Durme. 2014. Inferring User Political Preferences from Streaming Communications. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, ACL, pages 186–196.

Ulrike von Luxburg. 2007. A Tutorial on Spectral Clustering. *Statistics and Computing*, 17(4):395–416.

Junting Ye, Shuchu Han, Yifan Hu, Baris Coskun, Meizhu Liu, Hong Qin, and Steven Skiena. 2017. Nationality classification using name embeddings. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, CIKM, pages 1897–1906.