Less is more? How demographic sample weights can improve public opinion estimates based on Twitter data.

Pablo Barberá
University of Southern California
pbarbera@usc.edu

Abstract

Twitter data is widely acknowledged to hold great promise for the study of political behavior and public opinion. However, a key limitation in previous studies is the lack of information about the sociodemographic characteristics of individual users, which raises concerns about the validity of inferences based on this source of data. This paper addresses this challenge by employing supervised machine learning methods to estimate the age, gender, race, party affiliation, propensity to vote, and income of any Twitter user in the U.S. The training dataset for these classifiers was obtained by matching a large dataset of 1 billion geolocated Twitter messages with voting registration records and estimates of home values across 15 different states, resulting in a sample of nearly 250,000 Twitter users whose sociodemographic traits are known. To illustrate the value of this approach, I offer three applications that use information about the predicted demographic composition of a random sample of 500,000 U.S. Twitter users. First, I explore how attention to politics varies across demographics groups. Then, I apply multilevel regression and postratification methods to recover valid estimate of presidential and candidate approval that can serve as early indicators of public opinion changes and thus complement traditional surveys. Finally, I demonstrate the value of Twitter data to study questions that may suffer from social desirability bias.

Working paper. This version: September 21, 2016. I would like to thank John Ahlquist, Drew Dimmery, Pat Egan, Martin Elff, Mik Laver, Jeff Lewis, Thomas Leeper, Jonathan Nagler, Molly Roberts, Arthur Spirling, David Sontag for helpful comments and discussions. I also gratefully acknowledges financial support from the Gordon and Betty Moore Foundation and the Alfred P. Sloan Foundation.

Twitter data is widely acknowledged to hold great promise for the study of social and political behavior (Mejova, Weber and Macy, 2015; Jungherr, 2015). In a context of plummeting survey response rates, where particular subpopulations are harder to reach through traditional sampling methods, tweets represent public expressions of political opinions, which have been found to be correlated with offline opinions and behavior (O'Connor et al., 2010; DiGrazia et al., 2013; Vaccari et al., 2013), and thus do not suffer from some of the limitations of polls. In addition, Twitter allows researchers to unobtrusively observe individuals' behavior, from media consumption to social interactions, without the limitations inherent to self-reported answers in a survey. And this data collection process can take place in real-time and with a level of granularity that could only be achieved in the past at great coast.

Despite the great promise of this source of data, whether social media data can potentially replace or complement public opinion polling remains an open question (Schober et al., 2016). A key challenge that remains to be overcome is the lack of sociodemographic information about Twitter users. Unlike other social media platforms, Twitter does not require its users to provide basic information about themselves, such as gender or age. This restricts our ability to extrapolate any finding that relies on Twitter data to the entire population. The adjustments that are often conducted in survey research – i.e. the use of survey weights to recover the representativeness of a sample – are simply not possible in this context.

Beyond this methodological concern, the lack of individual-level covariates restricts the range of questions that can be studied with Twitter data. For example, if we are interested in measuring support for political candidates in a primary election, it is not currently possible to subset only those who are registered as members of that party. Being able to identify income, gender, and race would enable studies of social segregation in online settings. We could also study social inequalities in political behavior at a much more granular level if we were able to observe the individual traits of Twitter users.

The contribution of this paper is to address this key methodological problem in previous studies. I apply supervised learning methods to estimate the age, gender, race, party affiliation, propensity to vote, and income of any Twitter user in the U.S. This work improves upon previous studies on latent attribute inference based on Twitter data (Al Zamal, Liu and Ruths, 2012; Chen et al., 2015; Mislove et al., 2011; Pennacchiotti and Popescu, 2011; Rao et al., 2010) in two different ways. First, by relying on a ground truth dataset at least two orders of magnitude larger than those used in previous studies, this method achieves significantly better performance in this task. Second, and most importantly, the features used to predict Twitter users' latent traits are not "costly" – they can be measured using no more than 5 API calls per user. This makes it easy to scale to large datasets, and maximizes the applicability of this method to future studies.

This paper also provides an application of these methods to questions of substantive interest. In particular, I examine three different ways in which Twitter data can contribute to studies of public opinion. First, I examine how attention to political issues varies across sociodemographic groups using a panel of 500,000 users, randomly selected. This provides a new measure of issue salience that does not suffer from the limitations of the "most-important problem" question (Wlezien, 2005). Second, I show that measures of sentiment of the President and candidates for President based on Twitter data can serve as early indicators of changes in presidential and candidate approval. To adjust for sampling and selection biases in this dataset and recover its representativeness, I apply multilevel regression and post-stratification methods (Lax and Phillips, 2009; Park, Gelman and Bafumi, 2004; Wang et al., 2014). Finally, I examine how support for the Black Lives Matter movement varies across demographic groups and over time, as an example of how these methods can help in the study of hard-to-reach populations and topics that may suffer from social desirability bias.

Predicting Sociodemographic Traits with Twitter Data

Background and Related Work

Previous studies have approached the problem of estimating the sociodemographic characteristics of Twitter users using one of three approaches. One option is to apply supervised machine learning methods to a training dataset of users whose traits are known, usually by human coding. For example, Chen et al. (2015) used Amazon Mechanical Turk to label the ethnicity, gender, and age of 2,000 users, and then ran different classifiers using features from users' tweets, their neighbors, and their profile pictures. Pennacchiotti and Popescu (2011) employed a similar method with a sample of 6,000 users who stated their ethnicity in their descriptions, and 10,000 users who added themselves to a public directory of Democrats and Republicans on Twitter. Al Zamal, Liu and Ruths (2012) used the same source for political orientation, and 400 tweets from users announcing their own birthday to identify age. They considered a similar set of features – both information about users' tweets and about their friends and followers.

A second approach is to rely on indirect methods, such as extracting Twitter users' names, and comparing those with existing datasets with distributions of gender by first name and of ethnicity by last name to compute a probability of being male or female, and Caucasian, African-American, etc. (Mislove et al., 2011). A different type of indirect approach was used by Culotta, Ravi and Cutler (2015) – using website audience data, they show that followers of the Twitter accounts of these websites have a similar demographic composition. Within this category we would also find unsupervised methods that detect latent communities based on interactions on Twitter, building upon the assumption that behavior is homophilic (Conover et al., 2012; Barberá, 2015).

Finally, a few studies have taken advantage of recent developments in computer vision to predict the age, gender, and race of Twitter users based on their profile pictures (see e.g. An and Weber, 2016).

All three approaches have limitations that limit their scalability to large samples of users. Indirect methods do not perform well when sociodemographic traits are not heavily correlated with behavior, and name-based methods cannot be applied when new names are not included in the lists of names tagged by gender, which limits their applicability. For example, nearly 110,00 of 250,000 (44%) randomly selected U.S. Twitter users (see Applications section) did not report a first name that appears in the Social Security Administration baby names dataset (Blevins and Mullen, 2015). Supervised methods do not suffer from this problem but, because of their use of small, self-selected samples, they require collecting "costly" features in order to achieve high accuracy. Measuring features such as the text of a user's neighbors (her followers and those that she follows) is very time-consuming because it requires thousands of API calls, making this method impractical for any sample larger than a few hundre users. Methods based on pictures are easier to scale than machine learning methods applied to text or networks, but they can only be applied to visible traits, and they assume that users choose to display a picture of themselves, which may not be the case in a context like Twitter, where anonymity is allowed and even encouraged.

The aim of this paper is to develop a new approach that overcomes these limitations and allows any researcher to (1) estimate the age, gender, race/ethnicity, income, propensity to vote, and party affiliation of (2) any Twitter user, and (3) with fewer than 5 API calls per user.

Method

Even if the sociodemographic characteristics of Twitter users cannot be directly observed, there are at least three different types of information that researchers could use to infer them.

Text of users' tweets. A range of previous studies have shown significant differences in language use between men and women (Newman et al., 2008), liberals and conservatives (Sylwester and Purver, 2015), individuals of different age (Schwartz et al., 2013) and race groups (Florini, 2013). Language use indicates not only differences in personality or opinions, but also in interests and activities, which may also be correlated with users' sociodemographic characteristics. Text in microblogging platform such as Twitter often includes emoji characters – ideograms that include facial expressions, objects, flags, among others, and which often can convey more complex ideas than single words, and whose usage may also differ across demographic groups.

Users' friends. Previous studies have systematically found that the characteristics of users' friends – who they decide to follow – are highly correlated with their own characteristics (Chen et al., 2015; Al Zamal, Liu and Ruths, 2012). This result is consistent with the strong homophilic patterns commonly found in social networks (McPherson, Smith-Lovin and Cook, 2001). However, collecting information about the entire network of a given user is costly, often requiring multiple API calls. Instead, the approach I propose here is to focus on which *verified* accounts users decide to follow, and use this information to predict their latent traits.[1] If we consider Twitter as a news media (Kwak, Moon and Lee, 2012), these following decisions can also be informative about users' interests and preferences.

Profile information. Twitter allows users to write a 140-character description of themselves. Previous studies have shown that words contained in this field are highly predictive of sociodemographic traits (Pennacchiotti and Popescu, 2011). For example, many users write in this field that they are "husband/wife" of another user. We consider this information here as well although we note that its applicability is limited by the sparsity of the data, as many users do not fill the 'description' field.

Each of these sources of information provides an almost infinite number of potential features that could be used to train a machine learning classifier. Here, I estimate five different models, each of them with a different feature selection strategy. (K represents the total number of features.)

1. *Bag-of-words*: only words (unigrams) used by more than 1% and less than 90% of the users in the training dataset (K=34,092).

2. *Bag-of-emoji*: only emoji characters used by more than 1% and less than 90% of the

[1] Verification is granted by Twitter to public figures, including celebrities, media outlets, and politicians, in order to certify that their profile corresponds to their real identity. The full list of verified accounts is publicly available at `http://twitter.com/verified`.

users in the training dataset (K=627).

3. *Bag-of-followers*: of the over 154,000 accounts currently verified, only accounts with more than 10,000 followers and English or Spanish as their account language are included (K=61,659). Similar to an adjacency matrix or a document-feature matrix, the set of features for each individual will be a vector of length K with value 1 if the user follows that particular account and 0 otherwise.

4. *Bag-of-profile-words*: only words (unigrams) and emoji characters mentioned in the profile descriptions of more than 1% and less than 90% of the users in the training dataset (K=25,500).

5. *Combined classifier*: all features included in the four previous models.

Given the size and sparsity of each of these five feature matrices, I estimate supervised learning classifiers that perform well in this context. In particular, I use a logistic classifier with L2 regularization, also known as ridge regression (see e.g. Hastie, Tibshirani and Friedman, 2009), as implemented in the `LIBLINEAR` C/C++ library (Fan et al., 2008).[2] Ridge regression imposes a penalty on the size of the coefficients associated with each feature, which increases bias but reduces variance, and thus improves predictive accuracy. This model is particularly useful for this application, where most words or friends are not informative about latent traits, and thus by reducing their associated coefficients we minimize overfitting. The degree of regularization is driven by the penalty parameter, λ, which is chosen using cross-validation. In this particular application, a value of λ=100 was found to maximize cross-validated accuracy.

Data

Geolocated tweets

The first step in the data collection process was to construct a list of U.S. Twitter users whose location is known with county-level granularity. To do so, I collected a random sample of 1.1 billion geolocated tweets from around the world between July 2013 and May 2014. Of these, nearly 250 million tweets from 4.4 unique million users were sent from the contiguous United States. The pairs of coordinates (longitude and latitude) in each tweet was then used

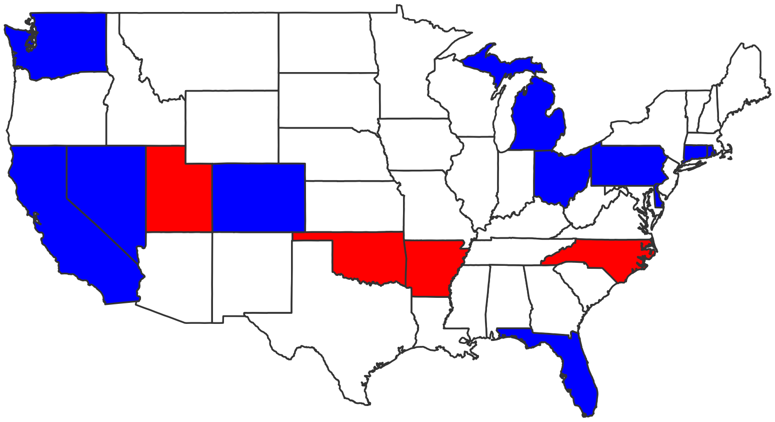
[2] This library is called from R using the `LiblineaR` package (Helleputte, 2013). All models were estimated using the New York University's High Performance Computing resources.

to identify the county and zipcode from which each of them was sent, using the shape files that indicate the polygons delimiting each of these geographical units. The "name" field in users' profiles was also extracted from all the tweets in this dataset, and parsed using regular expressions to split into first, middle, and last name. These two sources of information – geographic (county and zipcode) and name (first and last) – will be used to match Twitter accounts with their publicly available voting records. The resulting matches will be serve as training dataset for the supervised learning classifiers described in the previous section.

Voting Registration Records

The availability of voting registration records varies across states, depending on the rules imposed by their Secretaries of States. In around half of the cases, they are freely available upon request or after paying a small fee. These files generally contain the full name, residential address, party affiliation, gender, race, and past vote history for all voters that have ever registered to vote. In this project, I use voting records from 15 different states: Arkansas, California, Colorado, Connecticut, Delaware, Florida, Michigan, Nevada, North Carolina, Ohio, Oklahoma, Pennsylvania, Rhode Island, Utah, and Washington, as shown in Figure 1. While this set of states was chosen for convenience reasons (in all 15 states the voter records can be easily obtained online), it presents significant variation in electoral outcomes, population, and region. The voting records from each of these states was parsed and standardized to a common file format in order to facilitate the matching process. All the code necessary to run this step is available at `github.com/pablobarbera/voter-files`.

Figure 1: States included in the training dataset (in red states won by Romney in 2012; in blue states won by Obama in 2012)



Matching Process

A given Twitter account was matched with a voter only when there was a perfect and unique match of first name, last name, county, and state. In cases of multiple Twitter accounts or voters with identical first and last names in a county, they were matched at the zipcode level using the same method. This procedure is conservative on purpose – the goal is to create a training dataset with as little uncertainty as possible about users' true characteristics, in order to reduce measurement error. More sophisticated methods, based on geographic distance, could also be implemented in future work.[3] One limitation of this conservative approach is that it might induce bias in the classifier, as individuals included in the training dataset may not be representative of the entire population of registered voters.[4]

Table 1: Matching voting records and Twitter users.

State	Registered Voters	Twitter Users	Total Matches	%
Arkansas	1,582,012	32,372	4,615	14.2
California	17,811,391	554,213	65,079	11.7
Colorado	3,500,164	56,844	9,009	15.8
Connecticut	2,186,628	46,840	5,902	12.6
Delaware	645,329	13,008	1,923	14.8
Florida	13,037,192	260,604	36,308	13.9
Michigan	7,425,020	118,919	17,710	14.9
Nevada	1,438,967	57,069	6,724	11.8
North Carolina	5,413,637	127,463	14,292	9.5
Ohio	7,507,994	162,993	28,047	17.2
Oklahoma	1,983,727	48,780	6,746	13.9
Pennsylvania	8,231,634	168,873	21,537	12.7
Rhode Island	740,051	18,557	2,607	14.0
Utah	1,481,505	31,862	3,536	11.1
Washington	4,339,309	65,565	11,226	17.1
Total	77,324,560	1,763,962	233,132	13.2

Table 1 provides summary statistics for the sample sizes considered at each step. The

[3] Note that voters' residential address is available in all states; and these addresses could also be parsed to coordinates.

[4] For example, it is possible that individuals that belong to minority groups may be more likely to be matched to Twitter accounts, as their names are less common, and thus the number of duplicated matches is lower. I come back to this limitation later in the paper.

first column indicates the total number of registered voters in each state – their total sum correspond to between 35% and 50% of all registered voters in the U.S., depending on how these are defined. The second column shows the number of Twitter users in each state, based on the dataset of geolocated tweets. The third and fourth columns show the total number of Twitter users that were matched using this method, and the proportion that it represents over the total of Twitter users in each state. This proportion ranges from 9.5% in North Carolina to 17.2% in Ohio. While these proportions may seem low, Bond et al. (2012) were only able to match around 33% of Facebook users to voter records, despite having access to users' birthdates in a much less anonymous social networking site, where users are less likely to use pseudonym.

Since the residential address in which each voter is registered is also publicly available, this dataset can also be matched with home property records to obtain a rough estimate of each user's income. In particular, I queried the Zillow API for the 'zestimate' for each address – an estimate of the market value of each individual home, calculated for about 100 million homes in the U.S. based on public and user-submitted data points. More information is available at: `www.zillow.com/zestimate/`. This quantity is then normalized by multiplying it for the ratio of the median home value in each state over the median home value in the U.S., in order to have comparable values across different states. Despite this transformation, note that home values are still a noisy proxy for citizens' income. For example, I cannot distinguish whether the home is owned or rented. Despite these limitations, this variable provides a good estimate of the tercile of the income distribution each individual belongs to.

The final step in the data collection process was to download the list of 'friends' for all 233,132 users matched with voting records, as well as their 1,000 most recent tweets. Since 99% of the users in this sample follow fewer than 25,000 accounts, it is possible to construct the feature matrix with fewer than 5 API calls per user. (Each API call can return 200 tweets or 5,000 friends.) After excluding private and suspended Twitter accounts, the total size of the training dataset is 201,800 Twitter accounts.

Variables

After merging and cleaning all the datasets, in my analysis I will focus on six sociodemographic variables, recoded as follows:

- **Gender**: male or female.

- **Age**: 18-25, 26-40, 40+ (approximately three deciles of age distribution of Twitter users in the sample).
- **Race**: African-American, Hispanic/Latino, Asian/Other, White
- **Party**: Unaffiliated, Democrat, Republican.
- **Vote**: turnout in 2012 presidential election (yes/no).
- **Income**: normalized home value lower than $150,000, between $150,000 and $300,000, and greater than $300,000 (approximately three deciles of home value distribution in the sample).

Not all these variables are available in all states. Whenever the voter file does not contain one of these traits, they are marked as missing in the analysis and thus excluded.

Results

Tables 2 reports the performance of the classifiers for all sociodemographic characteristics. Figure 2 provides a graphical representation of the results to facilitate its comparison. In order to examine the performance of each model, I provide as a baseline the proportion of individuals in the modal categories for each variable (male, 40+, white, unaffiliated, voted in 2012, home value $150K-$300K), as well as the sample size included in the estimation. Accuracy was computed on a 20% random holdout sample. Note that the total sample size is lower than 201,800 in some cases because not all variables are available in some states, or for all individuals. For example, race is only available in Florida and North Carolina.

I find that the performance of the classifiers is in all cases better than random or choosing the modal category, with the exception of the bag-of-emoji and bag-of-profile-words models for some variables. When compared with previous studies, the levels of accuracy reported here are comparable or higher to those previously achieved. For example, Chen et al. (2015) achieve 79% accuracy for ethnicity, 88% accuracy for gender, and 67% accuracy for age. Al Zamal, Liu and Ruths (2012) obtain 80% accuracy for age, 80% accuracy for gender, and 92% accuracy for political orientation.[5]

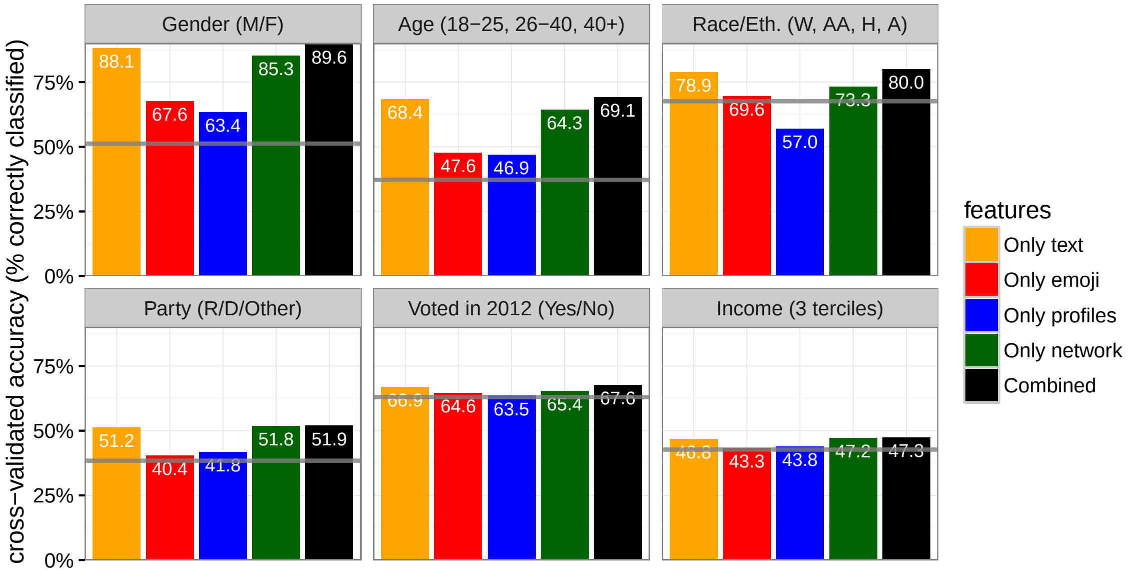
The performance of these classifiers varies across sociodemographic variables. Gender and age appear to be the easiest to predict, with the text-based classifier obtaining the highest accuracy, consistently with previous studies in computational linguistics that find large

[5] Note that these results are based on features that are much more costly to obtain, or use self-selected samples where it is easier to achieve good performance because they are easier to classify.

Table 2: Performance of machine learning classifiers (accuracy on 20% holdout dataset)

	Gend.	Age	Race	Party	Vote	Inc.
Baseline (mode)	51.2	37.2	67.6	38.4	63.0	42.7
N (users, 1000s)	130	202	40	174	196	159
Categories	2	3	4	3	2	3
Text classifiers						
Bag-of-words	88.1	68.4	78.9	51.2	66.9	46.8
Bag-of-emoji	67.6	47.6	69.6	40.4	64.6	43.3
Bag-of-profile-words	63.4	46.9	57.0	41.8	63.5	43.8
Network classifiers						
Bag-of-followers	85.3	64.3	73.3	51.8	65.4	47.2
Combined classifier						
All four feature groups	89.6	69.1	80.0	51.9	67.6	47.3

Figure 2: Performance of machine learning classifiers



differences in language use across gender and age groups. Accuracy is also high in the text-based classifier for race, although this result is not surprising, given that one of the largest minorities in the sample speaks a language other than English.[6] The party identification clas-

[6] At the same time this result also raises questions about the performance of the classifier across different groups within this ethnic community (e.g. first- vs second-generation immigrants). The use of this method implies in practice that members of this community are identified mostly based on their language, and depending on how it is going to be applied, it may lead to a problem of representativeness of the predicted sample of

sifier also appears to perform better than the baseline, but close to 50% of the sample is misclassified. One explanation for this result is that most users do not tweet about politics (as we show in the Applications section) or follow any political accounts, and thus there is simply no information to accurately predict their political preferences. Finally, the turnout and income classifiers barely perform better than the baseline, which could be due again to the difficulty of predicting interest in politics or the higher degree of measurement error in the income variable.

To further investigate the performance of these classifiers, Table 3 provides additional information about the two main models I will use in the remainder of this paper.[7] Here, I disaggregate each variable into individual categories, and compute accuracy, precision, and recall for each dichotomized indicator.[8] This analysis reveals some of the limitations of this approach. For example, recall is low for the Republican and Low Income categories, which means that these individuals are harder to distinguish from the rest of categories. This problem is particularly an issue for the Asian/Other category, although it can be explained by the very low sample size for this group; for this reason, this category is excluded from further analyses. Despite these weaknesses, these results do not show any imbalance across categories, and in all cases the accuracy of the classifiers is clearly higher than the baseline, which strongly justifies its use in the rest of this paper.

An alternative method to evaluate the performance of the classifiers is to identify the emoji characters, words and accounts with the highest and lowest estimated coefficients in the regularized logistic regression. Appendix A on page 31 details the results of this analysis, which show high face validity and are consistent with previous studies of language use in psychology and linguistics.

As discussed earlier, one key limitation of this approach is that the training dataset is not representative of Twitter users. Since it only contains individuals who report their real names on their profiles and who are registered to vote, it is likely that the dataset contains

Hispanics with respect to the entire population of Hispanics on Twitter. Additional evidence of this limitation of the model is the low recall levels reported in Table 3; in other words, many Hispanic Twitter users are not being identified as such, probably because they don't tweet in Spanish as often.

[7] As I explain in the Applications section, I choose to use the network-based model to predict the sociodemographic characteristics of the panel of Twitter users in order to avoid potential endogeneity issues due to the use of text in the prediction task and then also to measure outcomes of interest. For example, if the use of the word "obama" is a good predictor of being a Democrat, then it wouldn't be surprising to find that (predicted) Democrats mention Obama more often than Republicans. Using a network-based model avoids these potential issues

[8] Accuracy is computed as the proportion of correctly predicted values. Precision is the percentage of observations predicted to have value 1 that have true value 1. Recall is the percentage of observations with true value 1 that are correctly predicted to have value 1.

Table 3: Performance of machine learning classifiers, by category

	Network			All			
Variable	A	P	R	A	P	R	%
Gender							
Female	85	83	90	89	90	90	51.2
Age							
18-25	81	66	62	85	72	71	26.7
26-40	73	64	54	75	67	61	36.1
$\geq$ 40	74	63	76	78	70	76	37.1
Race/ethnicity							
African Am.	88	54	73	91	73	53	13.4
Hisp./Latino	84	52	65	89	71	54	17.2
Asian/Other	97	14	13	98	24	10	1.6
White	77	88	77	82	83	93	67.6
Party							
Democrat	67	54	55	66	53	58	38.4
Republican	77	57	31	75	50	40	36.3
Unaffiliated	60	48	62	63	52	54	25.2
Turnout							
Voted	65	67	89	68	70	85	63.0
Income							
Low	72	47	20	68	44	44	27.5
Middle	51	45	76	55	47	54	42.7
High	72	55	31	71	52	42	29.8

A = accuracy; P = precision; R = recall; % = prop.

users who are more active and more likely to use Twitter to consume political information. In order to evaluate how the classifiers perform out of sample, I took a random sample of 2,000 Twitter users in the U.S. (see next section for details on sample selection) and used the crowd-sourcing platform CrowdFlower to code their gender, race/ethnicity, and age based on their name and profile pictures. Table 4 shows that the out-of-sample performance of this models is lower than in-sample, as expected, but still significantly above any baseline classifier.[9]

[9] The sample size in these models is lower than 2,000 because users without a profile picture of a person were excluded, as well as individuals whose gender/age/ethnicity was not identifiable from their pictures.

Table 4: Out-of-sample performance.

Variable	Observed (%)	Network A	Network P	Network R	N
Gender					
Female	52	73	67	87	1,461
Race/ethnicity					
African Am.	14	89	83	27	1,203
Hisp./Latino	24	78	55	70	1,203
Asian/Other	0	96	0	NA	1,203
White	62	69	84	61	1,203
Age					
18-25	51	59	74	30	1,329
26-40	40	61	55	24	1,329
≥ 40	9	45	12	89	1,329

A = accuracy; P = precision; R = recall.

Applications

This section demonstrates the value of the methodological contribution introduced here by providing three applications to key substantive political science questions. All three applications rely on the same dataset: a panel of Twitter users whose sociodemographic characteristics are estimated using the machine learning classifiers described above. Tracking a fixed set of users has the main advantage of not requiring any particular set of keywords during the data collection process, which overcomes some of the limitations of getting access to Twitter data.[10] It can allow researchers to use prior user behaviors to detect their biases (Lin et al., 2013; Diaz et al., 2016), and provides information about when users in the minority do not share their opinion, thus controlling for 'spiral of silence' effects (Hampton et al., 2014). In addition, the availability of sociodemographic information about each user can be employed to correct for known differences between Twitter users and the target population using post-stratification (Little, 1993).

To construct the panel of Twitter users, I selected a random sample of 500,000 users in the United States. The random selection was achieved by sampling users based on their numeric ID: first, I generated random numbers between 1 and the highest numeric ID assigned at the

[10] Historical Twitter data is not available through the public API. As a result, if a new topic arises and researchers are interested in recovering old tweets mentioning a particular keyword, they would need to collect this data using expensive data reselling services, such as those provided by GNIP.

time (3.3 billion); then, for each number I checked whether the user existed, and whether the 'time_zone' field in their profile was one of the time zones in the United States or whether their 'location' field mentioned the full name or abbreviation of a U.S. state or one of the top 1,000 most populated cities; if the user met one of these conditions, she was included in the sample. The final step was to collect all of their tweets from January 1st, 2015 to August 1st, 2016 (a total of 227,154,772 tweets), and the accounts they follow.[11] Tweets were then stored in a Google BigQuery table to efficiently query this database (all queries required for the analyses in next section took less than 30 seconds).

After this dataset was collected, I applied the method above to predict the probability that each individual belongs to each political and sociodemographic category considered here.[12] Individuals who do not follow any verified accounts (11.7% of the sample) are excluded from the analysis. This random sample is predicted to be 49% male and 51% female; 20% ages 18-25, 36% ages 26-40, and 44% ages 40+; 44% Democrat, 31% Republican, and 25% unaffiliated; 26% low income, 44% medium, 30% high; 11% African-American, 17% Hispanic, 2% Asian/Other, 70% White; and with turnout of 63%. Table **??** in Appendix B provides summary statistics for the other characteristics of this sample of users (number of tweets, friends, and followers).

Measuring Issue Salience

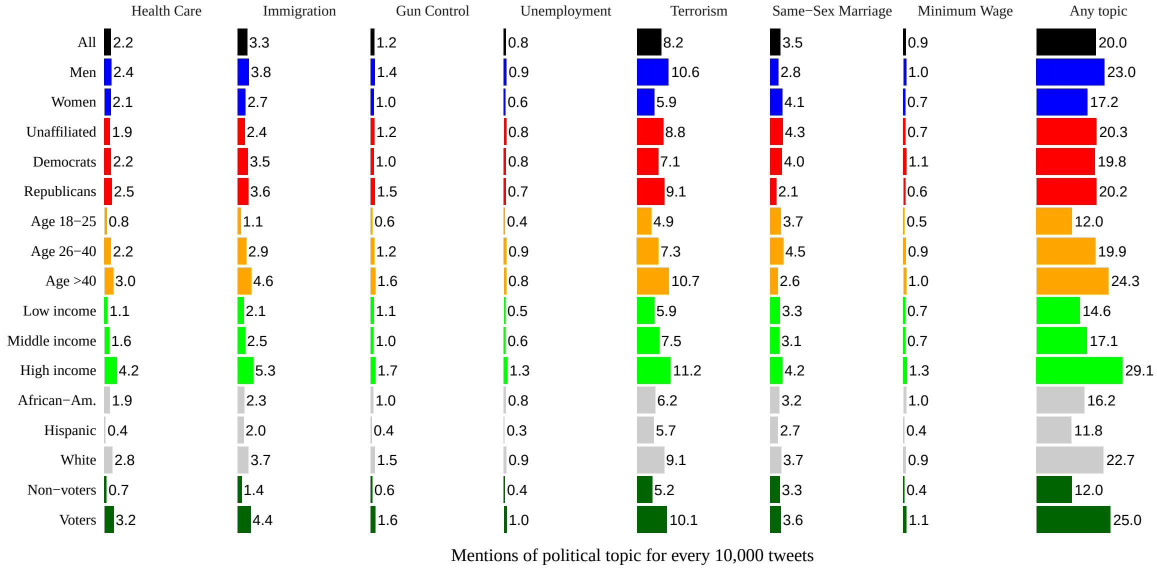
Studies of political responsiveness, party competition, and agenda-setting dynamics require granular measures of citizens' attention to political issues; i.e. issue *salience* (Behr and Iyengar, 1985; Berelson, Lazarsfeld and McPhee, 1954; Epstein and Segal, 2000). The most common approach – using survey responses to the "most important problem" question – suffers from well-known methodological limitations: it confuses importance of issues with the degree to which they are a problem (Wlezien, 2005), and responses are given in an artificial context, which leads to potential concerns about recall. In contrast, social media posts are spontaneous reactions to political news and events. This section provides a proof of concept of how this new source of data can complement traditional survey methods in the study of issue salience, by relying on a panel of Twitter users whose sociodemographic traits are known.

[11] The dataset does not include *all* the tweets ever sent by these accounts because Twitter only allows access to the 3,200 most recent tweets from a given account via the API, and some tweets by the most active accounts may have been excluded.

[12] I use the network-based method in order to avoid endogeneity concerns with the use of text to predict independent variables and as the outcome of interest.

Figure 3 displays the frequency with which different demographic groups discuss each of seven relevant political issues.[13] Since some groups tend to send more tweets than others, the counts here are normalized to number of tweets for every 10,000 tweets sent by each group.

Figure 3: Estimated attention to political issues, by sociodemographic group: mentions for each 10,000 tweets sent



As is commonly assumed, this Figure confirms that only a small proportion of Twitter users are interested in politics: only 0.2% of all tweets sent by this panel are related to any of the topics often considered to be central to political debate. In fact, 83% of this sample never mentioned *any* of these keywords in their tweets. When we compare across issues, terrorism is the most frequently mentioned, whereas unemployment and minimum wage (which, as part of economic issues, often appear as the most frequent responses to the "most important problem" question) are rarely discussed.

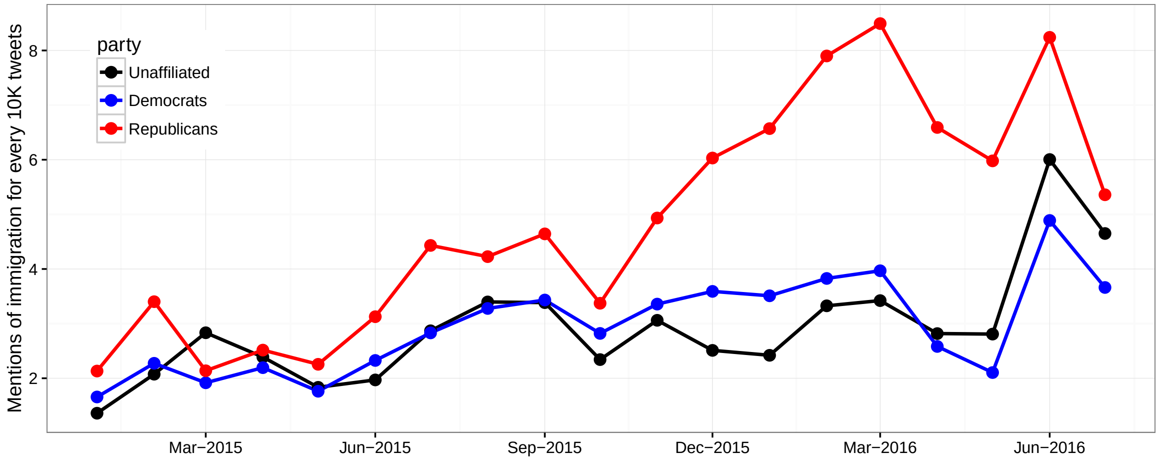
However, the most interesting patterns emerge when we examine different demographic groups. Most of these results are consistent with previous studies: men and individuals over 40 years old appear to discuss politics more often than women and younger individuals. (This result is replicated with other political topics, as shown in Table 10 in Appendix B.) The

[13] Tweets were classified into issues using a keyword-based method, ignoring case and punctuation. The set of keywords was: obamacare, health care, affordable care (health care); immigr* (immigration), gun control, second amendment, 2nd amendment, gun reform, gun rights (gun control), unemploy* (unemployment), terroris*, isis (terrorism), loveislove, equal marriage, same sex marriage, gay marriage, lovewins (same-sex marriage), minimum wage, minimumwage, raisethewage, fightfor15 (minimum wage).

only exception to this pattern is same-sex marriage. Income and propensity to vote are also positively correlated with higher interest in these political topics. The differences between Republicans and Democrats are not large, with the exception of two highly partisan topics: terrorism (discussed more often by Republicans) and same-sex marriage (discussed more often by Democrats).

Another key advantage of the panel design is that it allows us to track change in issue salience over time, as illustrated in Figure 4. Here, I provide evidence that is consistent with the "Trump hypothesis" – the claim that the candidacy of Donald Trump increased attention to the issue of immigration among Republican voters. By focusing on users who are predicted to be registered Republicans, I can indeed test whether that is the case. As shown here, mentions of immigration on Twitter increased after June 2015, and this change appears to be largely due to Republicans discussing this issue.[14]

Figure 4: Testing the "Trump hypothesis": did the salience of the immigration issue increase among Republicans during the primary election season?



Estimating Public Opinion with Twitter Data and Sociodemographic Weights

The increase in the use of Twitter for political purposes has led many researchers to examine whether specific patterns in the stream of tweets mirror offline public opinion, or if they might be even able to predict election outcomes (O'Connor et al., 2010; Tumasjan et al., 2010; DiGrazia et al., 2013). Despite this apparent success, different studies have demonstrated that the predictive power of tweets has been highly overstated (Gayo Avello, Metaxas

[14] Of course, this change could be due to a number of factors, not only Donald Trump's decision to run for president. Unfortunately, we cannot observe the counterfactual.

and Mustafaraj, 2011; Jungherr, Jürgens and Schoen, 2012; Beauchamp, 2014; Jungherr et al., 2016). The two most important limitations of previous research are the fact that sampling bias and self-selection bias are neglected: not all sociodemographic groups are equally present in Twitter, and some groups are much more likely to tweet about political topics (Barberá and Rivero, 2014). This is another application where the use of a panel design with known sociodemographic characteristics can represent an improvement upon previous approaches. Disaggregating across groups can provide a more granular view of public opinion that may help us understand why Twitter-based measures could be biased. Furthermore, demographic covariates can be used to adjust estimates of public opinion to their joint distribution in the population, and thus recover its representativeness.

As an empirical evaluation of these two innovations, I now turn to an analysis of tweets by the panel of Twitter users that mention President Obama, as well as the two candidates in the 2016 U.S. presidential election, Hillary Clinton and Donald Trump.[15] Each tweet was assigned a sentiment score on a scale from 0 to 1 using a supervised learning classifier.[16]

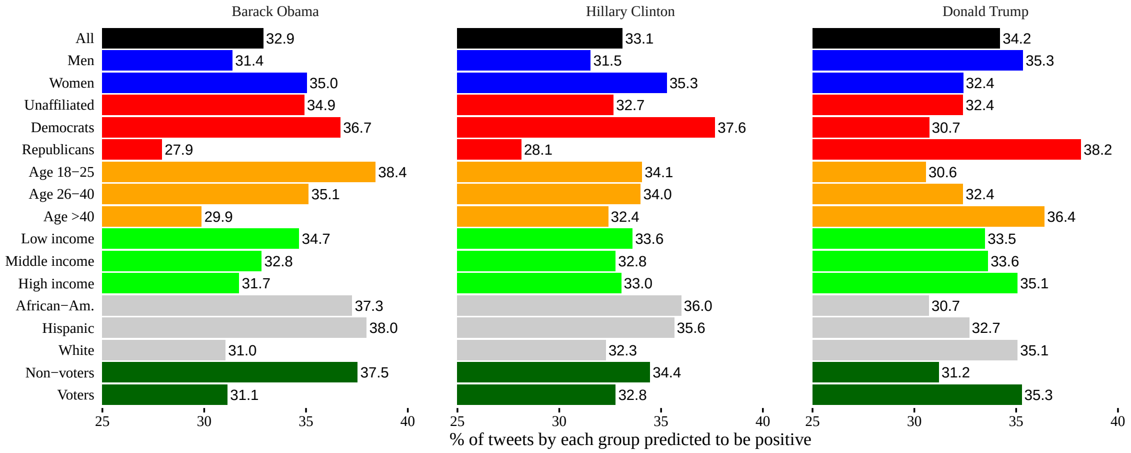
Figure 5 offers a first look at the results of this analysis. Here, I aggregate the sentiment scores computed for each tweet at the group level, in order to examine what individual characteristics predict expressions of support on Twitter. Consistently with public opinion polls, women, registered Democrats, young citizens, those in the lowest income tercile, as well as African-Americans and Hispanic users are much more likely to express support for Barack Obama and Hillary Clinton. In contrast, Donald Trump's supporters on Twitter appear to be disproportionally male, registered Republicans, over 40 years old, and white.

Once we know Twitter users' demographic characteristics, and how they are correlated with expressed support for political candidates, can we use this information to combine these estimates into a single measure of support that can be considered "representative" of the U.S. population? To examine this question, I adopt a similar approach as Wang et al. (2014), who were able to successfully forecast the 2012 Presidential election with highly

[15] Tweets related to each politician were extracted using keywords, ignoring case and punctuation: barack, obama (Barack Obama), hillary, clinton, imwithher, i'mwithher (Hillary Clinton), donald, trump, makeamericagreatagain, america great again (Donald Trump).

[16] Three different models were estimated, one for each candidate, using a regularized logistic classifier and choosing unigrams and bigrams that appear in 2 or more tweets as features, after standard pre-processing steps such as converting to lowercase and removing punctuation with the `quanteda` package (Benoit and Nulty, 2016). Each classifier was trained with a random sample of 5,000 tweets mentioning each politician, labeled by CrowdFlower workers (Benoit et al., 2016) in response to the question "does this tweet express support for [candidate name]?", with options "yes", "no", and "unclear/unrelated". Only "yes" and "no" answers are included in the training dataset. Despite the short length of individual tweets, all three classifiers achieved acceptable performance: accuracy on a 20% random holdout set was 72.3% for Obama, 74.1% for Clinton, and 71.0% for Trump.

Figure 5: Sentiment score in tweets mentioning Obama, by sociodemographic group



non-representative polls of Xbox users.

The intuition behind this method is that to improve the representativeness of a sample of Twitter users, one can weight each sociodemographic group according to the proportion of the population in that group. This method, commonly known in the survey research literature as *post-stratification* (Little, 1993), allows us to compensate for the fact that, for example, Republicans are underrepresented on Twitter. More in detail, this approach consists on partitioning the sample of Twitter users into J cells based on the combination of all sociodemographic characteristics, and then take a weighted sum of the average sentiment in each cell, $\hat{y}_j$, where each cell-level sentiment estimate is weighted according to the size N_j of that cell in the population:

$$\hat{y}_{PS} = \frac{\sum_{j=1}^{J} N_j \hat{y}_j}{\sum_{j=1}^{J} N_j} \tag{1}$$

This method will yield better results when the cells become more fine-grained, since the assumption of random sampling within each cell becomes more likely to hold; however, at the same time that also increases its sparsity, which can lead to noisy cell-level estimate. A common solution to this issue is to turn to a model-based strategy, multilevel regression and poststratification (MRP), which relies on a Bayesian hierarchical model to obtain better estimates for sparse cells (Lax and Phillips, 2009; Park, Gelman and Bafumi, 2004; Wang et al., 2014). For computational reasons, the model here is simpler: I fit a linear regression, where the dependent variable is the probability that each individual tweet expresses support, and the independent variables are a set of matrices that contain the probabilities that the individual who published it belongs to each of the sociodemographic groups considered:[17]

$$y_{ij} = \alpha + \beta_1 \text{gender}_j + \beta_\mathbf{2} \text{race}_j + \beta_3 \text{party}_j + \beta_4 \text{age}_j + \beta_5 \text{income}_j \qquad (2)$$

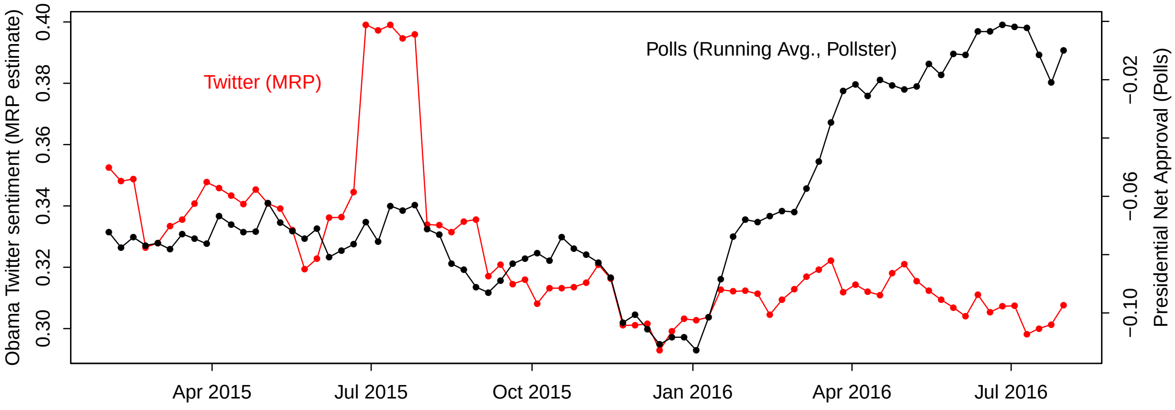
where gender_j, race_j, party_j, age_j, income_j are probability vectors for each individual j. After estimating the regression model, I compute the predicted proportion of supportive tweets in each cell, and then aggregate to the population level using data from the 2012 Congressional Cooperative Election Study (Ansolabehere and Schaffner, 2012), which includes all the relevant sociodemographic variables used here, and had a sample of over 50,000 respondents, large enough to give an accurate measure of the number of individuals in each cell in the general population.

Figure 6 shows the results of applying this method to the set of tweets published by users in the panel earlier described (in red; scale on left axis), compared with an average of polls about presidential approval compiled by Pollster. The quantity displayed here is *net presidential approval*. Both time series are displayed are the week level, with the estimates computed with tweets and polls from the last 30 days leading to each time point. As this figure shows, in 2015 there is a close correspondence between these two time series: the lowest and highest point of each series overlap in time, and changes over time appear to be correlated as well. The correlation between these two variables is $r = 0.70$ in 2015, but it becomes $r = -0.16$ for the entire time series; as the recent increase in approval according to polls does not appear to be reflected in the Twitter time series. As Figure 11 in Appendix B shows, this result is replicated if we disaggregate across party groups. The high sentiment scores for July 2015 are also a clear outlier, and appear to be due to the Supreme Court decision on same-sex marriage.

Although the results clearly demonstrate that, even after the adjustment, Twitter-based estimates of public opinion do not appear to be a good replacement of surveys, perhaps they could be used as an early indicator of changes. Figure 6 indeed indicates that changes on Twitter are taking place before they're registered in the polls, suggesting that individuals on Twitter are leading public opinion. Table 5 provides a more systematic test of the relationship between these these two variables using time series regression. Model 1 examines

[17] A key difference here with respect to Wang et al. (2014) is that all the variables in the model are probabilities, instead of the predicted values from the supervised learning models. This allows some of the uncertainty in the previous step of the estimation to propagate here.

Figure 6: Comparing Twitter- and survey-based measures of presidential job approval



whether contemporaneous values of adjusted Twitter sentiment are correlated with job approval based on surveys. Model 2 tests if this effect is robust to controlling for previous values of net approval in surveys. Finally, using an error-corrected model Model 3 evaluates whether lagged values and/or changes in net approval on Twitter predict current values of net approval according to surveys. Models 4 to 6 replicate this analysis but including only data prior to 2016.

Table 5: Time Series Regression

	(1)	(2)	(3)	(4)	(5)	(6)
Sentiment	-0.21	0.01		0.34∗∗	0.10∗∗	
(Twitter)	(0.14)	(0.03)		(0.05)	(0.04)	
Net approval		1.00∗∗	1.00∗∗		0.83∗∗	0.88∗∗
(Survey, Lagged)		(0.02)	(0.02)		(0.08)	(0.09)
Sentiment			0.15∗∗			0.16∗∗
(Twitter, Diff.)			(0.06)			(0.06)
Net approval			-0.01			0.07∗
(Twitter, Lagged)			(0.03)			(0.04)
Constant	0.00	-0.00	0.00	-0.19∗∗	-0.05∗∗	-0.04∗
	(0.05)	(0.01)	(0.01)	(0.02)	(0.02)	(0.02)
N	79	78	78	49	48	48
R2	0.03	0.96	0.96	0.49	0.85	0.86

Standard errors in parentheses. Signif.: ∗10% ∗∗5%. DV: Net presidential approval in surveys
Period of analysis: 01/01/2015–01/08/2016 (models 1–3) and –01/01/2016 (models 4–6).

The results considering the entire period show that changes on Twitter sentiment predict movement in presidential approval, even in a time series with such high serial correlation, as evidenced by the coefficient of the lag dependent variable 3. However, the overall degree of support on Twitter does not appear to be significantly and positively correlated with presidential approval, against expectations. This appears to be driven entirely by the recent increase in approval, as evidenced in Models 4 to 6, where the regressions are re-estimated including only data from 2015.

Figure 7: Sentiment score in tweets mentioning Clinton, and Trump, by sociodemographic group

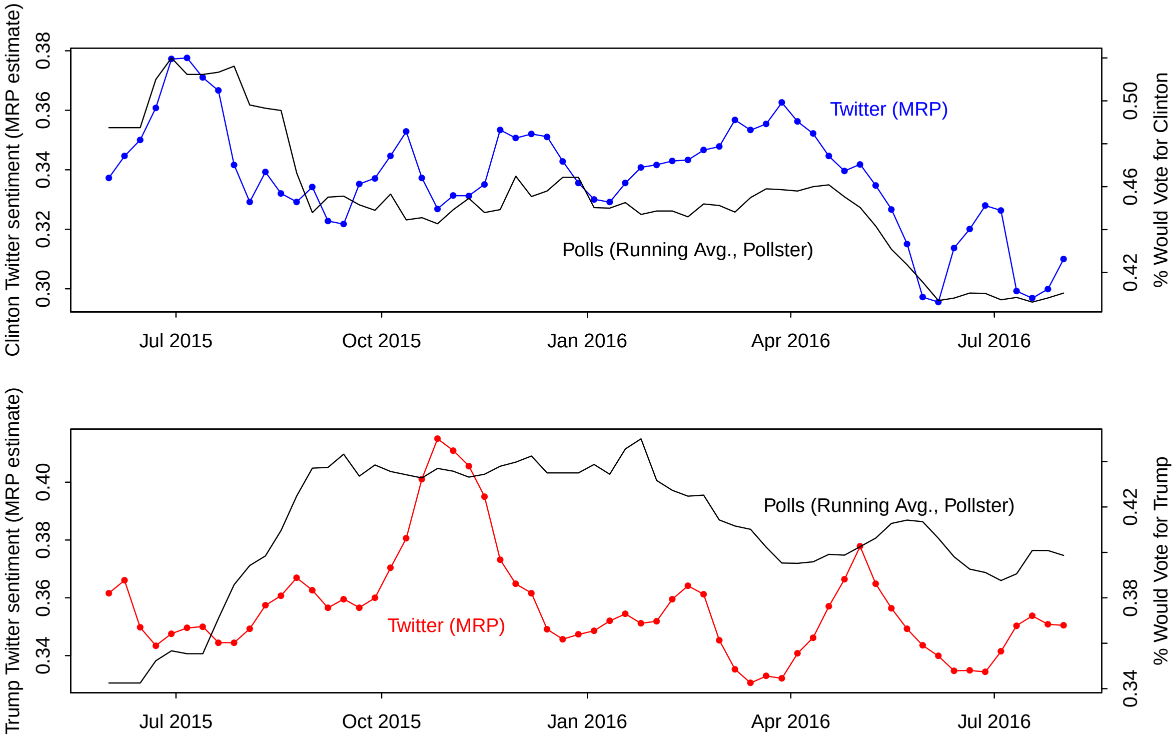


Figure 7 replicates the analysis above for tweets mentioning Hillary Clinton and Donald Trump, and comparing the MRP-adjusted estimate of support for each candidate with a polling average based on data from Pollster (in particular, a 30-day moving average based on general election Clinton vs. Trump match-ups). In both cases the two time series are positively correlated ($r = .76$ for Clinton and $r = .53$ for Trump), although for Trump there do appear to exist large gaps between Twitter-based estimates and public opinion polls.

Measuring Public Opinion About Potentially Sensitive Topics

Survey responses on sensitive topics such as drug use, racial prejudice, and attitudes towards corruption suffer from social desirability bias due to item nonresponse or under/overreporting (for a review, see e.g. Tourangeau and Yan, 2007). Changing the mode of administering the survey and the setting in which it is conducted, as well as the use of randomized response techniques (Warner, 1965) and list experiments (Blair and Imai, 2012; Glynn, 2013), are effective strategies to reduce this source of bias. However, the implementation of these techniques often increases the cost of conducting the survey and require large sample sizes in order to study differences across demographic groups. Here I explore whether the use of Twitter data could provide a new method to overcome this challenge.

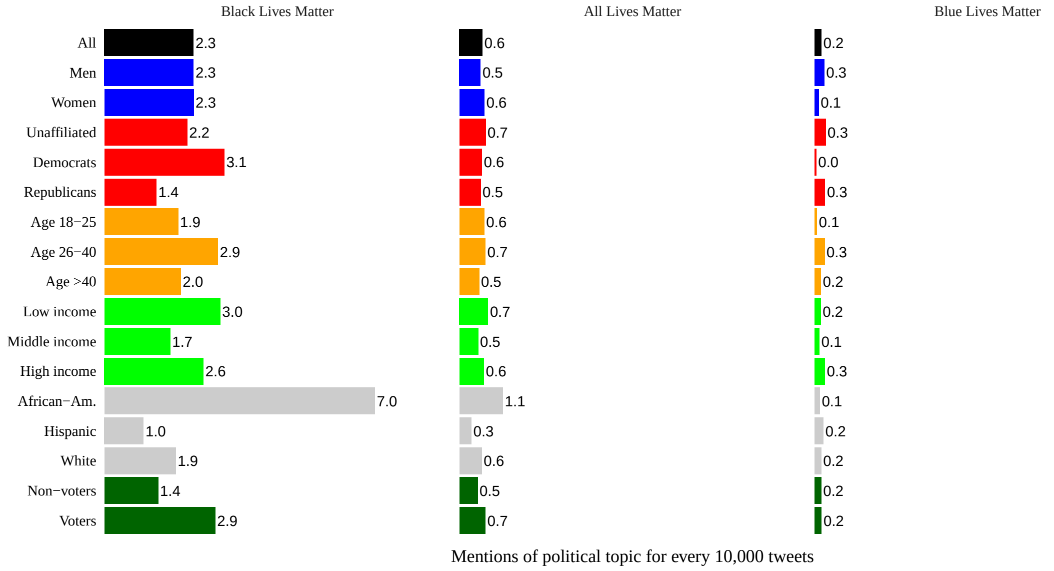
The case I will examine is support for the Black Lives Matter movement (BLM). Over the past couple of years, BLM has been able to successfully increase awareness about racial inequality and police brutality against the African-American community in the U.S. It gained prominence in the national media after the deaths of Michael Brown and Eric Garner in the summer of 2014, and the widespread protests that they ensued in Ferguson, Missouri and New York City (Freelon, McIlwain and Clark, 2016). Critics of this movement often respond to their claims by using the phrase “All Lives Matter”. More recently, also the phrase “Blue Lives Matter” has been used to imply support for the police.

Despite the importance of this movement in the current political context, few surveys have asked citizens about their opinions on this cause. A relevant exception is a survey conducted by the Pew Research Center in Feb-May of 2016, which found that support for Black Lives Matter was higher among African Americans (65%) than whites (40%), white Democrats (64%) than white Republicans (20%), and white adults ages 18–30 (60%) than ages 30–49 (26%) and ages 50–64 (37%). Another survey conducted in July 2015 by Rasmussen Reports found that 78% of likely voters thought “All Lives Matter” reflected better their point of view than “Black Lives Matter”.

Figure 8 offers an illustration of how the use of Twitter data could complement this set of surveys. Here, I show the frequency with which different sociodemographic groups mentioned all three versions of the “Lives Matter” slogans.[18] When compared with the results on Figure 3, I find that mentions of BLM are as frequent as mentions of health care or immigration, but less frequent than mentions of terrorism. Consistently with survey results,

[18] Tweets were selected using keywords: first, all tweets mentioning “livesmatter” or “lives matter” were extracted; then I subset only those that mentioned “black lives” or “blacklives” (Black Lives Matter), “all lives” or “allives” (All Lives Matter), and “blue lives” or “bluelives” (Blue Lives Matter)

Figure 8: Estimated attention to #BlackLivesMatter and related hashtags, by sociodemographic group: mentions for each 10,000 tweets sent

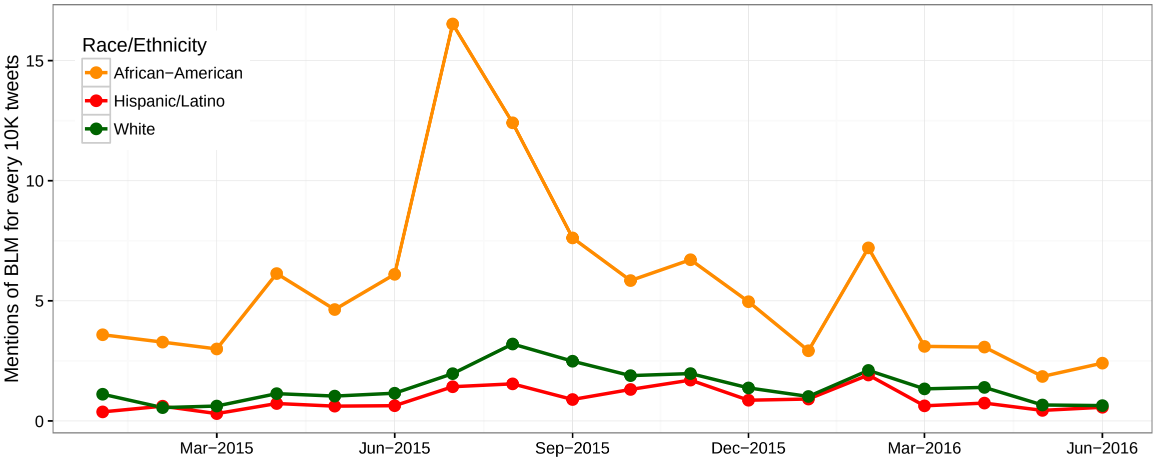


African Americans and Democrats are more likely to discuss this topic. These results are also similar to previous studies of the BLM on Twitter (Freelon, McIlwain and Clark, 2016; Olteanu, Weber and Gatica-Perez, 2015). In contrast, mentions of "All Lives Matter" are almost as frequent among Democrats as among Republicans; and the gap between African-Americans and whites is also much smaller. The differences are starker for "Blue Lives Matter": although this slogan is rarely mentioned, it is indeed the case that Republicans and white individuals are between 2 and 3 times more likely to use.

It is important to note that the use of any of these hashtags does not necessarily imply support (e.g. African-Americans might be mentioning "All Lives Matter" in order to explain how its use implicitly ignores structural racism), and future work could implement similar supervised learning classifiers as those used in the previous section.

Finally, Figure 9 tracks mentions of "Black Lives Matter" over time in order to examine to what extent the popularity of this movement grew at different rates across race/ethnicity groups. The graph shows a clear spike on July 2015, shortly after the shooting in a historically black church in Charleston, SC and the death of Sandra Bland, allegedly found hanged in a jail cell in Texas. Although attention to this topic increased for all groups, it appears to be the case that African-Americans reacted first, before attention to this issue become widespread across whites, as evidenced by the fact that the spike in attention in this second groups takes place on the following month.

Figure 9: Mentions of #BlackLivesMatter by race/ethnicity



Discussion

This article demonstrates that accurate estimates of the most relevant sociodemographic characteristics of Twitter users – those that are often used to recover the representativeness of survey respondents – can be accurately predicted from the text of their tweets and from who they decide to follow. I have shown the potential of this method by presenting three applications that demonstrate how Twitter-based estimates of public opinion can complement survey results: they may provide better measures of issue salience, serve as early indicator of changes in public opinion, and provide new insights into topics that may suffer from social desirability bias. While further work is needed, the work here highlights the still largely unexplored potential of social media data as a source of information to study social and political behavior.

Two main methodological challenges remain to overcome. First, how to improve out-of-sample performance? The classifier here was trained with data from users matched with voter registration records, which may not necessarily be representative of the population of Twitter users. A second concern is whether these methods are equally accurate for different subsets of each sociodemographic group. For example, if it's identifying Hispanics based on the use of the Spanish language, the group predicted to be Hispanic could actually be very different from the group of self-identified Hispanics in a survey. Perhaps one way to address this problem would be to run a survey of Twitter users, and examine whether their self-reports are similar to their predicted values. It may also be possible to conceive machine learning classifiers that calibrate the predicted values taking into account this problem.

References

Al Zamal, Faiyaz, Wendy Liu and Derek Ruths. 2012. Homophily and Latent Attribute Inference: Inferring Latent Attributes of Twitter Users from Neighbors. In *ICWSM*.

An, Jisun and Ingmar Weber. 2016. "# greysanatomy vs.# yankees: Demographics and Hashtag Use on Twitter." *arXiv preprint arXiv:1603.01973* .

Ansolabehere, Stephen and Brian Schaffner. 2012. "CCES common content, 2012." *Cooperative Congressional Election Study (distributor), version* 2.

Barberá, Pablo. 2015. "Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data." *Political Analysis* .

Barberá, Pablo and Gonzalo Rivero. 2014. "Understanding the political representativeness of Twitter users." Forthcoming in Social Science Computer Review.

Beauchamp, Nick. 2014. Predicting and Interpolating State-level Polling Using Twitter Textual Dat. In *APSA 2014 Anual Meeting Paper*.

Behr, Roy L. and Shanto Iyengar. 1985. "Television News, Real-World Cues, and Changes in the Public Agenda." *Public Opinion Quarterly* 49(1):38–57.

Benoit, Kenneth, Drew Conway, Benjamin E Lauderdale, Michael Laver and Slava Mikhaylov. 2016. "Crowd-sourced text analysis: reproducible and agile production of political data." *American Political Science Review* 110(2):XX–XX.

Benoit, Kenneth and Paul Nulty. 2016. *quanteda: Quantitative Analysis of Textual Data*. R package version 0.9.7-17.
URL: *https://CRAN.R-project.org/package=quanteda*

Berelson, Bernard R, Paul F Lazarsfeld and William N. McPhee. 1954. *Voting: A study of opinion formation in a presidential campaign*. University of Chicago Press.

Blair, Graeme and Kosuke Imai. 2012. "Statistical analysis of list experiments." *Political Analysis* 20(1):47–77.

Blevins, Cameron and Lincoln Mullen. 2015. "Jane, John ... Leslie? A Historical Method for Algorithmic Gender Prediction." *Digital Humanities Quarterly* .

Bond, R.M., C.J. Fariss, J.J. Jones, A.D.I. Kramer, C. Marlow, J.E. Settle and J.H. Fowler. 2012. "A 61-million-person experiment in social influence and political mobilization." *Nature* 489(7415):295–298.

Chen, Xin, Yu Wang, Eugene Agichtein and Fusheng Wang. 2015. A Comparative Study of Demographic Attribute Inference in Twitter. In *Ninth International AAAI Conference on Web and Social Media*.

Conover, Michael D, Bruno Gonçalves, Alessandro Flammini and Filippo Menczer. 2012. "Partisan Asymmetries in Online Political Activity." *EPJ Data Science* 1(1):1–19.

Culotta, Aron, Nirmal Kumar Ravi and Jennifer Cutler. 2015. Predicting the Demographics of Twitter Users from Website Traffic Data. In *Proceedings of the International Conference on Web and Social Media (ICWSM), in press. Menlo Park, California: AAAI Press.*

Diaz, Fernando, Michael Gamon, Jake M Hofman, Emre Kıcıman and David Rothschild. 2016. "Online and social media data as an imperfect continuous panel survey." *PloS one* 11(1):e0145406.

DiGrazia, Joseph, Karissa McKelvey, Johan Bollen and Fabio Rojas. 2013. "More tweets, more votes: Social media as a quantitative indicator of political behavior.".

Epstein, Lee and Jeffrey A. Segal. 2000. "Measuring Issue Salience." *American Journal of Political Science* 44(1):66–83.

Fan, Rong-En, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang and Chih-Jen Lin. 2008. "LIBLINEAR: A library for large linear classification." *Journal of machine learning research* 9(Aug):1871–1874.

Florini, Sarah. 2013. "Tweets, tweeps, and signifyin': Communication and cultural performance on "Black Twitter"." *Television &amp; New Media* p. 1527476413480247.

Freelon, Deen Goodwin, Charlton D McIlwain and Meredith D Clark. 2016. "Beyond the hashtags:# Ferguson,# Blacklivesmatter, and the online struggle for offline justice." *Available at SSRN* .

Gayo Avello, Daniel, Panagiotis T Metaxas and Eni Mustafaraj. 2011. Limits of electoral predictions using twitter. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*. Association for the Advancement of Artificial Intelligence.

Glynn, Adam N. 2013. "What can we learn with statistical truth serum? Design and analysis of the list experiment." *Public Opinion Quarterly* 77(S1):159–172.

Hampton, Keith, Lee Rainie, Weixu Lu, Maria Dwyer, Inyoung Shin and Kristen Purcell. 2014. "Social media and the 'spiral of silence'." *Pew Research Center, Washington, DC pewinternet. org/2014/08/26/social-mediaand-the-spiral-of-silence* .

Hastie, Trevor, Robert Tibshirani and Jerome Friedman. 2009. *The elements of statistical learning*. Springer.

Helleputte, T. 2013. "LiblineaR: linear predictive models based on the liblinear C/C++ library." *R package version* pp. 1–80.

Jungherr, Andreas. 2015. *Analyzing Political Communication with Digital Trace Data*. Springer.

Jungherr, Andreas, Harald Schoen, Oliver Posegga and Pascal Jürgens. 2016. "Digital Trace Data in the Study of Public Opinion: An Indicator of Attention Toward Politics Rather Than Political Support." *Social Science Computer Review* .
URL: *http://ssc.sagepub.com/content/early/2016/02/15/0894439316631043.abstract*

Jungherr, Andreas, Pascal Jürgens and Harald Schoen. 2012. "Why the pirate party won the german election of 2009 or the trouble with predictions: A response to tumasjan, a., sprenger, to, sander, pg, & welpe, im "predicting elections with twitter: What 140 characters reveal about political sentiment"." *Social Science Computer Review* 30(2):229–234.

Kwak, Haewoon, Sue B Moon and Wonjae Lee. 2012. More of a Receiver Than a Giver: Why Do People Unfollow in Twitter? In *ICWSM*.

Lax, Jeffrey R and Justin H Phillips. 2009. "How should we estimate public opinion in the states?" *American Journal of Political Science* 53(1):107–121.

Lin, Yu-Ru, Drew Margolin, Brian Keegan and David Lazer. 2013. Voices of victory: A computational focus group framework for tracking opinion shift in real time. In *Proceedings of the 22nd international conference on World Wide Web*. International World Wide Web Conferences Steering Committee pp. 737–748.

Little, Roderick JA. 1993. "Post-stratification: a modeler's perspective." *Journal of the American Statistical Association* 88(423):1001–1012.

McPherson, M., L. Smith-Lovin and J.M. Cook. 2001. "Birds of a feather: Homophily in social networks." *Annual review of sociology* pp. 415–444.

Mejova, Yelena, Ingmar Weber and Michael W Macy. 2015. *Twitter: A Digital Socioscope*. Cambridge University Press.

Mislove, Alan, Sune Lehmann, Yong-Yeol Ahn, Jukka-Pekka Onnela and J Niels Rosenquist. 2011. "Understanding the Demographics of Twitter Users." *ICWSM* 11:5th.

Newman, Matthew L, Carla J Groom, Lori D Handelman and James W Pennebaker. 2008. "Gender differences in language use: An analysis of 14,000 text samples." *Discourse Processes* 45(3):211–236.

O'Connor, Brendan, 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):1–2.

Olteanu, Alexandra, Ingmar Weber and Daniel Gatica-Perez. 2015. "Characterizing the demographics behind the# blacklivesmatter movement." *arXiv preprint arXiv:1512.05671* .

Park, David K, Andrew Gelman and Joseph Bafumi. 2004. "Bayesian multilevel estimation with poststratification: state-level estimates from national polls." *Political Analysis* 12(4):375–385.

Pennacchiotti, Marco and Ana-Maria Popescu. 2011. "A Machine Learning Approach to Twitter User Classification." *ICWSM* 11:281–288.

Rao, Delip, 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 pp. 37–44.

Schober, Michael F, Josh Pasek, Lauren Guggenheim, Cliff Lampe and Frederick G Conrad. 2016. "Research synthesis Social media analyses for social measurement." *Public Opinion Quarterly* p. nfv048.

Schwartz, H Andrew, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman et al. 2013. "Personality, gender, and age in the language of social media: The open-vocabulary approach." *PloS one* 8(9):e73791.

Sylwester, Karolina and Matthew Purver. 2015. "Twitter language use reflects psychological differences between Democrats and Republicans." *PloS one* 10(9):e0137422.

Tourangeau, Roger and Ting Yan. 2007. "Sensitive questions in surveys." *Psychological bulletin* 133(5):859.

Tumasjan, Andranik, Timm Oliver Sprenger, Philipp G Sandner and Isabell M Welpe. 2010. "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment." *ICWSM* 10:178–185.

Vaccari, Cristian, Augusto Valeriani, Pablo Barberá, Richard Bonneau, John T Jost, Jonathan Nagler and Joshua Tucker. 2013. "Social media and political communication. A survey of Twitter users during the 2013 Italian general election." *Rivista italiana di scienza politica* 43(3):381–410.

Wang, Wei, David Rothschild, Sharad Goel and Andrew Gelman. 2014. "Forecasting elections with non-representative polls." *International Journal of Forecasting* .

Warner, Stanley L. 1965. "Randomized response: A survey technique for eliminating evasive answer bias." *Journal of the American Statistical Association* 60(309):63–69.

Wlezien, Christopher. 2005. "On the salience of political issues: The problem with 'most important problem'." *Electoral Studies* 24:555–579.

A Top predictive features associated with each category

An alternative method to evaluate the performance of the classifiers is to identify the emoji characters, words and accounts with the highest and lowest estimated coefficients in the regularized logistic regression. Table 6 reports these sets of words and accounts.

To facilitate the interpretation, the coefficients in the network model were weighted by the number of followers (for accounts) and the models based on words and emoji characters were re-estimated using TF-IDF normalization.TF-IDF (Term Frequency Inverse Document Frequency) normalization calculates a value for each word in a document as an inverse proportion of the frequency of that word in a particular document relative to the percentage of documents that word appears in. As a result, rare and very frequent words receive lower weight. The performance of models estimated with TF-IDF normalization instead of word frequency is slightly lower in this particular application, but it facilitates the interpretability of the results.

The results on Table 6 have high face validity and are consistent with previous studies of language use in psychology and linguistics – see Schwartz et al. (2013) for a review. For example, females use more emotion words and mention psychological and social processes, whereas males use profane words and object references more often. Regarding age, the results show a pattern of progression in individuals' life cycle: from school and college, to work, and then to family (e.g. some of the most predictive words of being older than 40 are words related to children and grandchildren); and from an emphasis on expressing emotions, to more action and object references. Another strong sign that the method is correctly classifying individuals' race and ethnicity is that one of the best predictor of each category is the skin tone modifier, which change the aspect of face emojis.

Regarding party identification, it appears the use of words and emoji related to marriage equality (e.g. the rainbow emoji), reproductive rights ("women") and skin tone modifiers a good predictor of a Twitter user being affiliated with the Democratic party, reflecting the sociodemographic composition of this group. Republicans, on the other hand, appear to be more likely to discuss their faith on Twitter. Individuals with no party affiliation are likely to use words that are unrelated to politics. Although the results are not as good for the turnout classifier, words such as "vote" and "news" and the check emoji appear as the best predictors of having voted in 2012, as well profile words that indicate that an individual was not eligible to vote in 2012 ("19", "university", "18", "17", "16", etc.)

Finally, the emoji and words associated with different income levels indicate another limitation of this method: many of these refer to geographic locations where home values are generally low or high (e.g. fresno and sacramento vs san francisco or miami). However, most of these words indicate the models are capturing some signal: e.g. tweeting about flights, travel, and activites like gold of ski are good predictors of having high levels of income; as well as references to college degrees (“university”, “alum”, “lawyer”) or jobs in tech or other white-collar professions (“ceo”, “co-founder”, “software”, “digital”, “engineer”).

The results for the network-based model are also consistent with previous work and popular conceptions of the audience for each of these accounts. For gender, just like Culotta, Ravi and Cutler (2015), I find that following Ellen Degeneres is an excellent predictor of a Twitter user being female, whereas following SportsCenter and other famous sports figures is a good indicator of an user being male. Republicans and Democrats also follow accounts that align with their political preferences: Barack Obama, Rachel Maddow, Bill Clinton; and Fox News, Mitt Romney, and Tim Tebow, respectively. African Americans and Hispanics appear to be likely to follow popular figures in their community, such as Kevin Hart, Oprah Winfrey or LeBron James, or Pitbull, Jennifer Lopez, and Shakira. Whites, on the other hand, are more likely to follow country stars like Blake Shelton. Finally, following Miley Cyrus, UberFacts or Daniel Tosh is a good predictor of being younger than 25 years old, whereas following CNN, Oprah or Jimmy Fallon is more likely among users older than 40.

Table 6: Top predictive features (emoji, words, accounts, words in profiles) most associated with each category.

Female	💞, 👯, 💗, 👭, 💁, 💜, 👙, 😩, 💃, ♡, 😍, 😘, 🍹, 🌻, 🐶, ❤️, 🙅, 🙅. . .
	love, women, hair, girl, husband, mom, omg, cute, excited, <3, girls, yay, happy, hubby, boyfriend, :), can’t, baby, wine, thank, heart, nails. . .
	@TheEllenShow, @khloekardashian, @MileyCyrus, @Starbucks, @jtimberlake, @VictoriasSecret, @WomensHealthMag, @channingtatum. . .
	PROFILE: mom, mother, girl, mommy, lover, actress, ✨, love, wife :),❤️, 💜, alumna, mama, daughter, woman, sister, lady, yoga, ☀️, makeup. . .
Male	👬, 🔥, 💯, 💀, 😎, 🚀, 💩, 💨, ✊, 😏, 🔊, 🏿, 👨, 💦, 😈, 💰, 🎮. . .
	bro, man, wife, good, causewereguys, gay, great, dude, f*ck, nice, game, iphone, ni**a, church, time, #gay, girlfriend, bruh, sportscenter. . .
	@SportsCenter, @danieltosh, @MensHealthMag, @AdamSchefter, @ConanOBrien, @KingJames, @katyperry, @ActuallyNPH. . .
	PROFILE: father, guy, dad, husband, man, sports, engineer, developer, pastor, technology, musician, gamer, actor, dude, baseball, software, tech. . .
Age 18-25	👌, 💁, 😑, 😍, 😭, 😏, 😔, 😂, 🔫, 🙋, 😒, 😅, 🎓, 😘, ☺️, 🍻, 🐊. . .
	class, college, semester, life, (:, sportscenter, campus, best, literally, like, haha, just, :d, finals, classes, okay, professor, exam, studying. . .
	@SportsCenter, @wizkhalifa, @MileyCyrus, @danieltosh, @instagram, @EmWatson, @KevinHart4real, @UberFacts, @vine. . .
	PROFILE: university, major, ❤️, college, student, 16, future, fsu, class, ✨, ☀️, ●, state, ucf, snapchat, taken, ✌️, 17, 15, ig, 19, ♡, ❤️, (:, 14. . .
Age: 26–40	👰, 👶, 👯, 📷, 💪, 😕, ✈️, 😉, 💩, 😆, 🚲, 😝, 💅, 👎, 👊. . .
	excited, work, amazing, bar, awesome, wedding, #tbt, pretty, #nofilter, ppl, bday, time, lil, #love, yay, #latergram, office, game, tonight, boo, super. . .
	@danieltosh, @ConanOBrien, @jtimberlake, @StephenAtHome, @chelseahandler, @KimKardashian, @instagram, @NPR, @britneyspears. . .
	PROFILE: nerd, alum, designer, enthusiast, beer, sports, mommy, lover, gamer, engineer, husband, manager, co-founder, awesome, opinions, girl. . .
Age: ≥ 40	🎂, 😀, 🏁, 😇, 🐾, ➡️, ⚾, 💝, 🌹, 🌟, 🙏, ⭐, 🎸, 👣, 🏈. . .
	great, daughter, son, nice, r, good, ok, kids, congratulations, obama, hi, nbcthevoice, wow, happy, hope, beautiful, sorry, rock, grandson, amen. . .
	@jimmyfallon, @cnnbrk, @YouTube, @Pink, @TheEllenShow, @NBCTheVoice, @SteveMartinToGo, @Oprah, @sethmeyers, @FoxNews. . .
	PROFILE: retired, mom, author, grandmother, dad, kids, mother, conservative, father, children, estate, fan, family, professor, consultant, realtor. . .
Afric. Amer.	🏾, 💯, 👀, 😩, 👈, 🏿, 🙏, 👣, 😒, 👉, 👋, 💋, 🏃, 💨, 🙌, 🙅, 🏽. . .
	black, smh, #scandal, lol, god, iamsteveharvey, bout, yall, man, ni**a, morning, blessed, wit, y’all, lil, yo, bruh, lord, good, . . .
	@BarackObama, @instagram, @KevinHart4real, @Oprah, @KingJames, @stephenasmith, @LilTunechi, @Lakers, @YouTube, @MariahCarey. . .
	PROFILE: god, follow, famu, black, man, woman, ask, im, facebook, n, laid, ig, coach, kik, haitian, producer, women, pray, blessed, speaker. . .
Hisp./ Latino	▶️, 🇪🇸, ⚽, 😦, 🏽, 🎊, 😌, 💆, 😓, 💭, 💪, 👫, 😋, 🗽, 😱, 🔊, 🇲🇽 . . .
	miami, lmao, colombia, que, en, #miami, fiu, lmfao, hola, el, fiu, cuban, la, hialeah, hispanic, lol, :d, lmfaooo, tu. . .
	@instagram, @nytimes, @JLo, @ladygaga, @SofiaVergara, @KimKardashian, @shakira, @georgelopez, @justinbieber, @pitbull, @DJPaulyD. . .
	PROFILE: y, miami, en, la, que, fiu, puerto, el, mi, ón, colombian, dominican, es, mia, ig, del, venezolana, los, mexican, vida. . .

Note: Each row indicates the top 15-20 emoji/words/accounts that better predict each category, not the most common.

Table 6: Top predictive features (emoji, words, accounts, words in profiles) most associated with each category.

Asian/ Other	😭, 🏀, 🏽, 🍴, ☺️, 👌, 🙀, 📷, 🍣, 😝, 💀, 😓, 😴, 🐼, 😑, 🔥, 🆗, 👳, 🇰🇷... asian, ig, ucf, enthusiast, vietnamese, snapchat, line, food, step, uf, tweets, technology, story, university, earth, sushi, forget, finish, graduate, engineer ... @TheEllenShow, @cnnbrk, @azizansari, @BarackObama, @DalaiLama, @NBA, @mindykaling, @mashable, @UberFacts, @JLin7 ... PROFILE: y, miami, en, la, que, fiu, puerto, el, mi, ón, colombian, dominican, es, mia, ig, del, venezolana, los, mexican, vida...
White	🏁, ⛳, 🏻, ⚾, 🐘, 🇺🇸, 💁, ⚡, 😷, 🐊, 🐯, 🍀, 🐏, 😳, ♡, 🏼, 😅, 🌊 ... tonight, sweet, florida, ya, beach, blakeshelton, cat, haha, beer, think, night, asheville, great, baseball, dog, today, sure, lake ... @ActuallyNPH, @TheEllenShow, @blakeshelton, @jimmyfallon, @tomhanks, @danieltosh, @Pink, @FoxNews, @RyanSeacrest ... PROFILE: fan, mom, teacher, wife, husband, conservative, beach, beer, cat, retired, fsu, southern, country, nascar, married, unc, dogs, florida, nc, writer ...
Dem.	🏾, 👀, 😩, 🌈, →, 🏽, 🍸, ✨, 🍷, 💋, 🌹, 💯, 🐧, 💀, 👏, 💃, 💦, 🎬, 🇲🇽, 💅 ... philly, barackobama, la, sf, pittsburgh, women, nytimes, philadelphia, smh, president, gop, black, hillaryclinton, gay, republicans ... @BarackObama, @rihanna, @maddow, @billclinton, @khloekardashian, @billmaher, @Oprah, @KevinHart4real, @algore, @MichelleObama ... PROFILE: philly, activist, writer, liberal, pittsburgh, producer, los, philadelphia, sf, politics, democrat, advocate, angeles, actress, professor, ...
Rep.	🐊, 🇺🇸, 🏁, ⛳, ⚾, 😳, ❌, ➡️, 🏈, 🐘, ♡, ☀️, ❄️, 👸, ⚡, 🔴, ⭐, ⚡, 💛, ☕ ... foxnews, #tcot, church, christmas, oklahoma, florida, obama, great, realdonaldtrump, golf, beach, megynkelly, tulsa, byu, seanhannity ... @FoxNews, @danieltosh, @TimTebow, @MittRomney, @taylorswift13, @jimmyfallon, @RyanSeacrest, @Starbucks, @JimGaffigan ... PROFILE: conservative, jesus, wife, christian, florida, pastor, follower, husband, oklahoma, church, christ, god, married, fsu, grace...
Other	🐗, 🅾️, 🍻, 😷, 🌰, 💁, 😑, 🚬, 💸, ☀️, ✊, 💥, ❄️, 🍺, ✌️, 🐪, 🚀 ... ohio, arkansas, columbus, cleveland, cincinnati, utah, toledo, cavs, #wps, browns, ar, akron, hogs, bengals, kent, dayton, #cbj, reds ... @instagram, @SportsCenter, @KingJames, @vine, @AnnaKendrick47, @wizkhalifa, @WhatTheFFacts, @galifianakisz, @ActuallyNPH... PROFILE: ohio, cincinnati, cleveland, arkansas, columbus, cle, bgsu, kent, osu, university, ksu, buckeye, utah, im, akron, 19, 18, uc, toledo ...
Voted	🎂, ⚾, 🏾, →, 😀, ☔, 🍷, 😉, ▶️, 🐾, ☑️, 🐗, ™, 😜, ➡️, 🏈, 🎤, ☕, 🏻, ⭐ ... great, obama, san, did, kids, vote, cleveland, daughter, nc, news, disneyland, barackobama, church, romney, county, president, california... @BarackObama, @TheEllenShow, @jimmyfallon, @FoxNews, @azizansari, @blakeshelton, @MittRomney, @Starbucks, @RyanSeacrest... PROFILE: fan, teacher, mom, retired, husband, wife, director, &, estate, author, father, conservative, married, public, lover, educator, liberal ...
Did not vote	😅, 👌, 🎓, 😏, 😂, 💯, 🔫, 💏, 😍, ✊, 🆒, 🎧, 💀, ✌️, 💁, 😌, 💕 ... college, life, philly, pittsburgh, bro, sportscenter, miss, florida, im, sh*t, penn, ya, f*ck, gonna, guys, can't, man, actually, wanna ... @SportsCenter, @vine, @justinbieber, @wizkhalifa, @MileyCyrus, @UberFacts, @Eminem, @KendallJenner, @Jenna_Marbles... PROFILE: 19, university, 18, 17, ❤️, im, snapchat, class, major, ucf, ●, fsu, taken, penn, ig, ❤️, sc, 16, insta, fl...

Note: Each row indicates the top 15-20 emoji/words/accounts that better predict each category, not the most common.

Table 6: Top predictive features (emoji, words, accounts, words in profiles) most associated with each category.

Income: Low	💯, 😩, 😒, 💏, 👫, ❤️❤️, 😴, 😠, 💋, 👶, 🎂, ♡♡, 👈, ♡❤️, 💵, ✊, 😆, 👀, 😤, 🎶 . . .
	fresno, sacramento, bakersfield, work, lol, spokane, watching, good, ass, follow, wwe, #raw, :-), baby, wwe, need, im, ready, tired, sleep, bored . . .
	@instagram, @WhiteHouse, @YouTube, @ArianaGrande, @tomhanks, @stephenasmith, @KevinHart4real, @aliciakeys, @carmeloanthony . . .
	PROFILE: fresno, single, follow, old, im, god, married, mommy, mother, young, y, ❤️, sacramento, gamer, u, taken, i'm, loving, worker, black. . .
Income: Middle	👴, 😭, 🍔, 💄, 🚗, 😂, 🚂, 😊, 🐾, 🏁, 🎹, 🌱, ⛄, 🎡, 💔, ⚾, 😳, 🌀 . . .
	diego, denver, disneyland, vegas, utah, #sandiego, church, sd, tonight, disney, anaheim, las, colorado, worship, abc7, kings, lakewood, awesome. . .
	@vine, @Usher, @ZooeyDeschanel, @RyanSeacrest, @AdamSchefter, @rihanna, @rainnwilson, @robkardashian, @andersoncooper . . .
	PROFILE: beer, girl, colorado, pastor, wife, ❤️, vegas, nevada, dumb, laughter, student, diego, sdsu, simple, animal, jesus, web, ✌️, fitness, usf. . .
Income: High	✈️, 👯, 🎉, ▶️, ✈️, 🇫🇷, 🌴, ⛳, 🎿, 🐯, 🇬🇧, →, 🐗, ☀️, 💃, ❤️, 🍸, 🏆, ❄️ . . .
	sf, francisco, best, miami, class, san, great, thanks, nyc, la, congrats, beach, data, michigan, college, philly, flight, actually, #sf, nytimes, seattle . . .
	@cnnbrk, @jimmykimmel, @StephenAtHome, @adamlevine, @jimmyfallon, @TechCrunch, @neiltyson, @SteveMartinToGo, @nytimes. . .
	PROFILE: ceo, founder, university, sf, product, francisco, co-founder, software, alum, attorney, digital, lawyer, opinions, design, marketing, engineer. . .

Note: Each row indicates the top 15-20 emoji/words/accounts that better predict each category, not the most common.

A Additional tables and figures

Table 7: Twitter panel: summary statistics.

	Tweets		Friends		Followers	
	avg	med	avg	med	avg	med
All	496	29	376	153	812	71
Men	465	30	397	152	929	65
Women	525	28	356	154	701	78
Age 18-25	926	162	395	229	543	177
Age 26-40	382	27	373	175	927	72
Age >40	390	15	369	101	843	43
African-Am.	681	47	652	276	1147	146
Hispanic	528	31	404	222	932	78
Asian/Other	484	33	390	222	1425	62
White	458	26	323	121	714	62
Democrats	464	22	355	141	796	64
Republicans	416	20	343	111	557	54
Unaffiliated	651	73	453	232	1156	113
Non-voters	688	64	435	223	927	121
Voters	384	19	341	116	745	52
Low income	564	36	465	216	605	90
Middle income	482	21	301	111	666	55
High income	457	38	408	169	1211	81

Figure 10: Estimated attention to political and non-political topics, by sociodemographic group: mentions for each 10,000 tweets sent

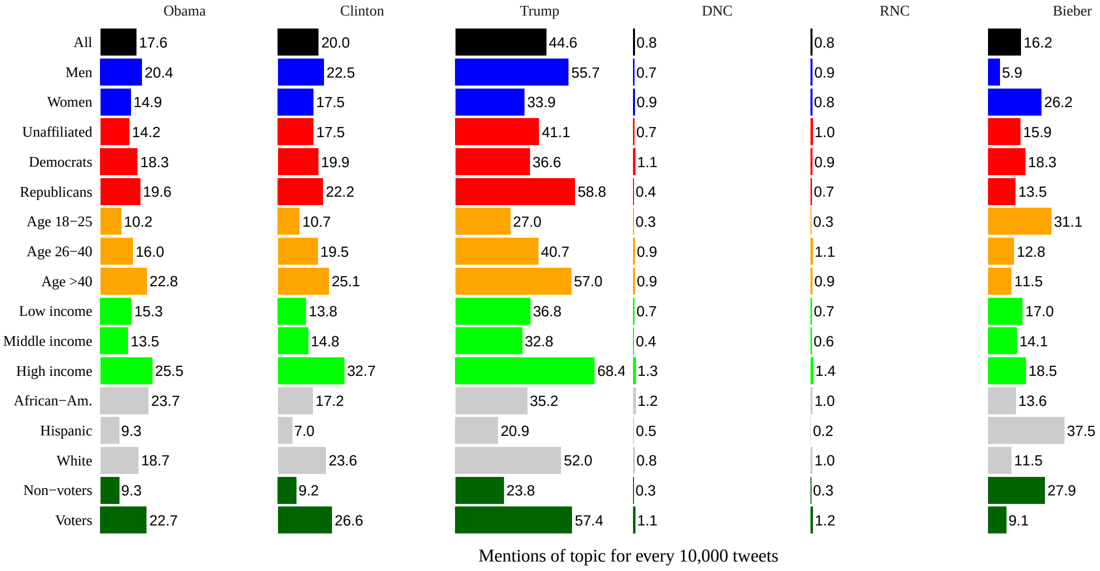
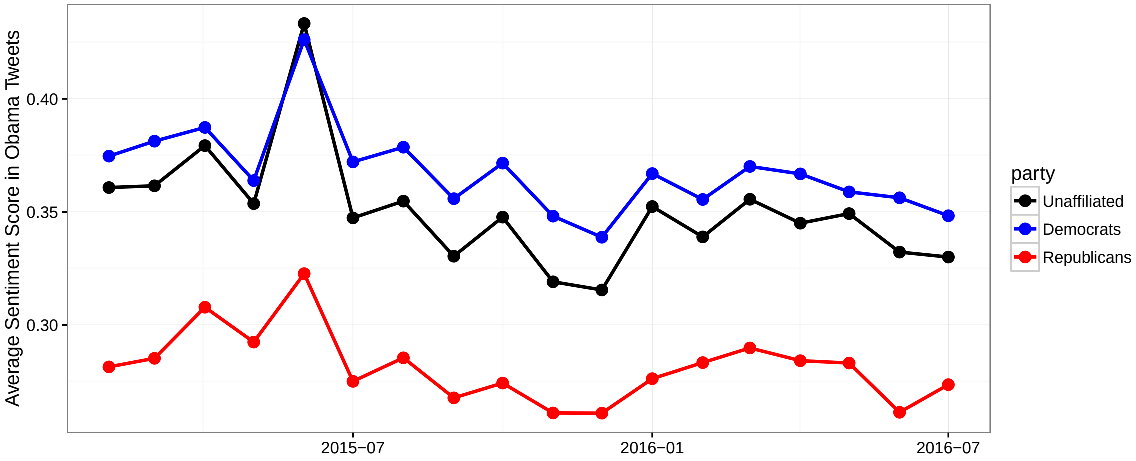


Figure 11: Average entiment score in tweets mentioning Obama, by month and party


Average Sentiment Score in Obama Tweets
party
Unaffiliated
Democrats
Republicans