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

Pablo Barberá
Center for Data Science
New York University
www.pablobarbera.com

Abstract

An important limitation in previous studies of political behavior using Twitter data is the lack of information about the sociodemographic characteristics of individual users. This paper addresses this challenge by developing new machine learning methods that will allow researchers to estimate the age, gender, race, party affiliation, propensity to vote, and income of any Twitter user in the U.S. with high accuracy. The training dataset for these classifiers was obtained by matching a massive 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. I illustrate the value of these new methods with two applications. First, I explore how attention to different candidates in the 2016 presidential primary election varies across demographic groups within a panel of randomly selected Twitter users. I argue that these covariates can be used to adjust estimates of sentiment towards political actors based on Twitter data, and provide a proof of concept using presidential approval. Second, I examine whether social media can reduce inequalities in potential exposure to political messages. In particular, I show that retweets (a proxy for inadvertent exposure) have a large equalizing effect in access to information.

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, tweets represent unfiltered 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). More generally, Twitter data also allows researchers to easily and unobtrusively observe social interactions in real-time, and to measure consumption of political information with a level of granularity that could only be achieved in the past at great cost.

Despite the great promise of this source of data, an important 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. As a result, researchers interested in working with Twitter data cannot construct survey weights to recover the representativeness of their samples in the same way that survey researchers combine probability sampling with post-stratification weights to reduce sampling selection bias (Schober et al. 2016).

Beyond this methodological concern, the availability of individual-level covariates would expand the range of questions that can be studied with Twitter data. For example, if we were interested in measuring support for political candidates in a primary election, it would allow us to subset only those that are affiliated with 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 develop new 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 can be measured using no more than 5 API calls per user, which makes it easy to scale to large datasets.

This paper also provides two applications of these methods to questions of substantive interest. First, I examine how attention to different candidates in the 2016 presidential primary election varies across sociodemographic groups using a panel of 200,000 users, randomly selected. This panel design overcomes some of the difficulties inherent to self-selection bias and, in combination with the sampling weights that can now be computed using the latent traits estimated with the method introduced here, could potentially allow researchers to recover the representativeness of estimates based on Twitter data. Second, I examine how exposure to

Working paper. This version: February 24, 2015. The author gratefully acknowledges financial support from the Gordon and Betty Moore Foundation and the Alfred P. Sloan Foundation.

political information on Twitter varies across sociodemographic groups. Merging data about who retweeted particular political messages with the lists of who each individual in this panel of 200,000 users follows, I am able to quantify direct (via following) and indirect exposure (via retweets). This analysis shows that even if direct exposure is highly unequal across social groups, the differences are significantly reduced once inadvertent exposure is considered. This result highlights the potential of social media to reduce inequalities in access to political information.

Background and Related Work

Previous studies have approached the problem of estimating the sociodemographic characteristics of Twitter users using one of two 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, Cheng (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 (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 (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 *et al* (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 2015a).

Both approaches have limitations that make it difficult to scale these methods to large samples of users. Indirect methods do not perform well when sociodemographic traits that 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 hundreds of API calls, making this method impractical for any sample larger than a few thousand users.

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 two 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. To test whether language use predicts users' latent traits, I will estimate two models, of increasing complexity: first, a logistic classifier with Elastic Net regularization (Zou and Hastie 2005) using only emoji characters as features (*bag-of-emoji*); second, a logistic classifier with Elastic Net regularization and Stochastic Gradient Descent (SGD) learning using word counts as features (*bag-of-words*, *BOW*), and applying a TF-IDF transformation. To reduce the size of the feature matrix, I will only consider emoji and words used by more than 1% and less than 90% of the users in the training dataset (627 emoji characters, 34,092 unigrams).

Users' friends. Previous studies have systematically found that the characteristics of users' neighbors – 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. Of the over 154,000 accounts currently verified, I select only 61,659 accounts

[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`.

with more than 10,000 followers and English or Spanish as their account language. Similar to an adjacency matrix, the set of features for each individual will be a vector of length 61,659 with value 1 if the user follows that particular account and 0 otherwise (*bag-of-followers*). As in the previous case, I will also estimate a logistic classifier with Elastic Net regularization and SGD learning to predict users' traits.

These two are not the only possible sources of information about users' characteristics. Twitter allows users to write a 140-character description of themselves in the profile, and this text has been used in previous studies to build training datasets (Pennacchiotti and Popescu 2011). As discussed in the previous section, first and last names also contain information about individuals' gender and ethnicity. However, even if in some cases these methods could lead to more accurate predictions, they are limited by the sparsity of the data: many users do fill the 'description' field or report a name contained in the existing name datasets.

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 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.

Voting Registration Records

The availability of voting registration records varies across states, depending on the rules imposed by their Secretaries of States. In most 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. 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.

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

All the code necessary to run this step is available at `github.com/pablobarbera/voter-files`.

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. More sophisticated methods, based on geographic distance, could also be implemented in future work. Note that voters' residential address is available in all states; and these addresses could be easily parsed to coordinates.

Table 1 provides summary statistics for the sample sizes considered at each step. The 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 a given citizens' wealth.

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.
- **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).

Results

Tables 2 reports the performance of the classifiers for all sociodemographic characteristics. 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 using 5-fold cross-validation. Note that the total sample size is lower than 40,000 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. In Table 3, I provide additional information about the performance of the two main classifiers, after disaggregating each variable into individual categories, and computing accuracy, precision, and recall for each dichotomized indicator.

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 models. 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 (2012) obtain 80% accuracy for age, 80% accuracy for gender, and 92% accuracy for political orientation. However, 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.

When comparing the two different methods, a clear pattern emerges: text-based features are as good or even better than network-based measures. The differences are particu-

Table 2: Performance of machine learning classifiers (Cross-validated accuracy, 5 folds)

	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-emoji	69.2	52.0	68.9	40.3	65.3	43.2
Bag-of-words	84.9	65.5	77.3	50.3	67.2	48.1
Network classifiers						
Bag-of-followers	85.3	63.1	77.6	50.7	64.2	45.7
Combined classifier						
Boe + Bow + Bof	88.7	68.3	80.5	53.9	67.6	49.4

Table 3: Performance of machine learning classifiers, by category

	Text			Network			
Variable	A	P	R	A	P	R	%
Gender							
Female	88	90	87	86	85	88	48.8
Age							
18-25	**85**	**72**	**68**	82	66	61	26.7
26-40	**74**	**67**	**43**	72	64	53	36.1
$\geq$ 40	**74**	**63**	**78**	73	62	76	37.1
Race/ethnicity							
African Am.	90	75	34	**89**	**89**	**30**	13.4
Hisp./Latino	**88**	**79**	**37**	86	78	25	17.2
Asian/Other	98	90	11	**98**	**75**	**2**	1.6
White	**77**	**77**	**97**	75	74	98	67.6
Party							
Democrat	62	51	55	**66**	**54**	**53**	38.4
Republican	76	55	22	**76**	**59**	**25**	36.3
Unaffiliated	**61**	**50**	**55**	58	47	66	25.2
Turnout							
Voted	**67**	**69**	**88**	65	66	91	63.0
Income							
Low	**73**	**50**	**18**	72	48	19	27.5
Middle	51	46	77	**50**	**45**	**79**	42.7
High	72	54	31	**72**	**55**	**27**	29.8

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

larly large for age, propensity to vote, and income. While there are differences in across these groups in who they follow (as evidenced by the fact that bag-of-followers features are also good predictors), it appears language traits are more indicative, which is consistent with previous research in computational linguistics. In the case of race, this results is not surprising, given that one of the largest minorities in the sample speaks a language other than English. However, at the same 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 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 Hispanics with respect to the entire population of Hispanics on Twitter. Additional evidence of this limitation of the model is the low recall levels of some of the classifiers; in other words, many Hispanic Twitter users are not being identified as such, probably because they don't tweet in Spanish as often. While this problem is perhaps more obvious in this case, it appears to apply to some other sociodemographic groups, such as Republican supporters.

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 4 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), in order to make them comparable to the TF-IDF normalization of the emoji and word-based models. These results 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. 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.

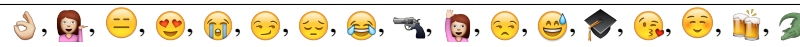
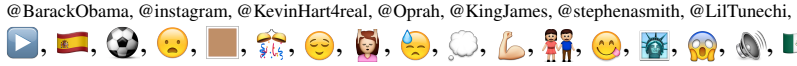
The results for the network-based model are also consistent with previous work and popular conceptions of the audience for each of these accounts. For gender, just like Culotta *et al* (2015), I find that following Ellen Degeneres is an excellent predictor of a Twitter user being female, whereas following 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.

One 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 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 5 shows that the out-of-sample performance of this models is lower than in-sample, as expected, but still significantly above any baseline classifier.

Table 5: Out-of-sample performance.

	Observed	Network			
Variable	(%)	A	P	R	N
Gender					
Female	52	75	69	86	1,461
Race/ethnicity					
African Am.	14	89	83	27	1,203
Hisp./Latino	24	81	76	31	1,203
Asian/Other	0	NA	NA	NA	1,203
White	62	71	69	95	1,203
Age					
18-25	51	63	66	57	1,329
26-40	40	63	56	36	1,329
≥ 40	9	72	17	63	1,329

Table 4: Top predictive features (emoji, words, accounts) 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...
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...
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...
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...
Age: $\geq$ 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...
African Am.	 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...
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...
Asian/Other	 i'm, asian, miami, lol, jacksonville, haha, tampa, :d, :), indian, allah, gainesville, like, india, orlando, #heatnation, #tampa, studying ... @TheEllenShow, @cnnbrk, @azizansari, @BarackObama, @DalaiLama, @NBA, @mindykaling, @mashable, @UberFacts, @JLin7 ...
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 ...
Democrat	 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 ...
Republican	 foxnews, #tcot, church, christmas, oklahoma, florida, obama, great, realdonaldtrump, golf, beach, megynkelly, tulsa, byu, seanhannity ... @FoxNews, @danieltosh, @TimTebow, @MittRomney, @taylorswift13, @jimmyfallon, @RyanSeacrest, @Starbucks, @JimGaffigan ...
Unaffiliated	 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...
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...
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...
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 ...
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 ...
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...

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

Applications

Estimating Public Opinion with Social Media 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, 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).

One solution to these two challenges is to analyze a panel of Twitter users whose sociodemographic characteristics are estimated with the methods I introduce here. Tracking a fixed set of users over time can allow researchers to use prior user behaviors to detect their biases (Lin et al. 2013; Diaz et al. 2016), and also can provide 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). A similar approach was used by Wang *et al* (2014) to forecast the 2012 Presidential election with highly non-representative polls of Xbox users.

As an empirical evaluation of the promise of these two methodological innovations, I now turn to an analysis of tweets by a panel of Twitter users that mention each of the most popular candidates in the 2016 Democratic and Republican election campaign, as well as President Obama. The goal is to learn about how often different sociodemographic groups mention, and apply sentiment analysis techniques in combination with a post-stratification adjustment in order to evaluate whether Twitter-based metrics approximate can public opinion polls on presidential approval and support for presidential candidates.

To construct the panel of Twitter users, I selected a random sample of 200,000 users in the United States. The random selection was achieved by sampling users based on their numeric ID: first, I generated random number between 1 and the highest numeric ID assigned at the 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 and the accounts they follow: a total of over 200 million tweets and 89 million accounts followed.[2]

After this dataset was collected, I applied the method above to predict the political sociodemographic characteristic of each user.[3] This random sample is predicted to be 47% male and 53% female; 43% ages 18-25, 24% ages 26-40, and 32% ages 40+; 35% Democrat, 32% Republican, and 32% unaffiliated; 30% low income, 26% medium, 44% high; 17% African-American, 19% Hispanic, 3% Asian/Other, 61% White; and with turnout of 61%. Figure 6 provides summary statistics for the other characteristics of this sample of users. Finally, I used Hadoop/MapReduce to extract the tweets that mention the names of the current U.S. president ("barack", "obama"), as well as the candidates in the Democratic and Republican presidential primary election, which includes the main hashtags of their campaigns (e.g. "Make America Great Again", "I'm With Her", "Feel The Bern")

Table 6: Twitter panel: summary statistics.

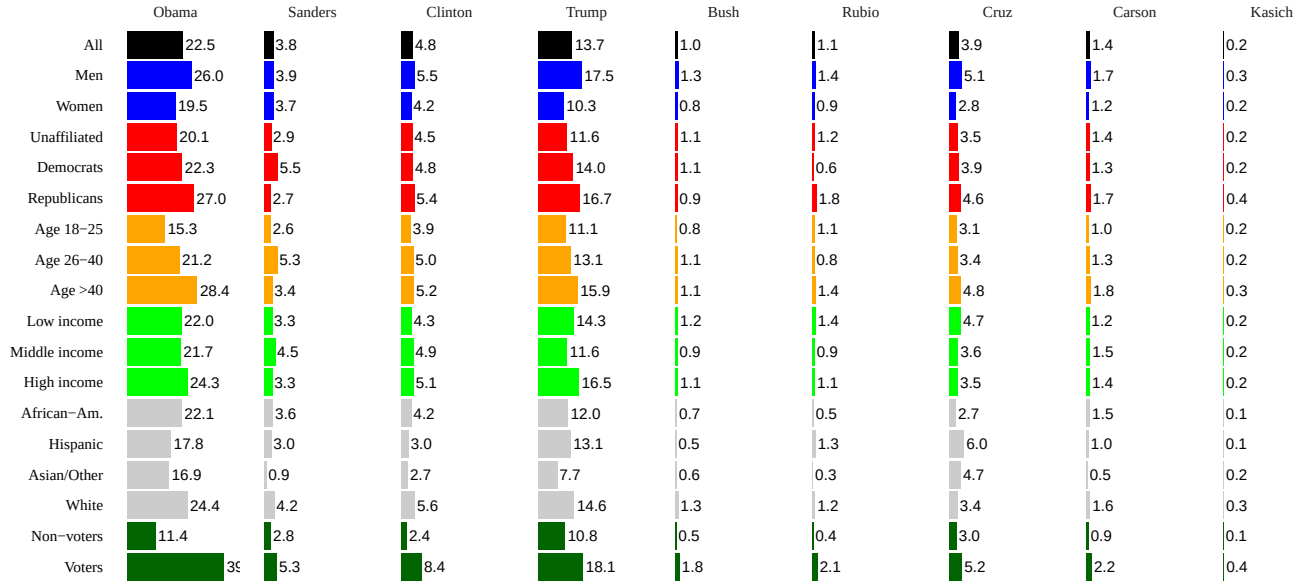
	Tweets		Friends		Followers	
	avg	med	avg	med	avg	med
All	3031	217	347	133	668	62
Men	2868	221	393	148	824	65
Women	3177	213	305	119	527	60
Age 18-25	3975	293	269	106	476	71
Age 26-40	2393	245	393	175	824	66
Age >40	2244	144	416	134	808	51
African-Am.	5660	573	535	219	925	128
Hispanic	2901	245	339	168	843	65
Asian/Other	3158	190	415	199	380	53
White	2325	166	293	100	555	51
Democrats	3422	337	375	167	711	82
Republicans	3393	246	395	161	783	69
Unaffiliated	2241	118	266	74	499	40
Non-voters	4146	510	414	205	592	110
Voters	2328	131	304	93	716	43
Low income	3794	285	398	163	623	70
Middle income	3367	335	407	177	716	80
High income	2309	136	276	89	670	48

Figure 1 displays estimates of how often these political actors are mentioned, measured as the number of tweets for each 10,000 tweets sent by each group. As is commonly assumed, only a small proportion of Twitter users are interested in politics: less than 0.2% of tweets mention Barack Obama and, in fact, less than 25% of users in the sample have ever mentioned the President. Among the presidential candidates, Donald Trump appears to be most popular, although of course most of the messages mentioning his name

[2] 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. This should be enough to cover at least one year of tweets for the large majority of accounts.

[3] 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 1: Estimated attention to politicians, by sociodemographic group: mentions for each 10,000 tweets sent

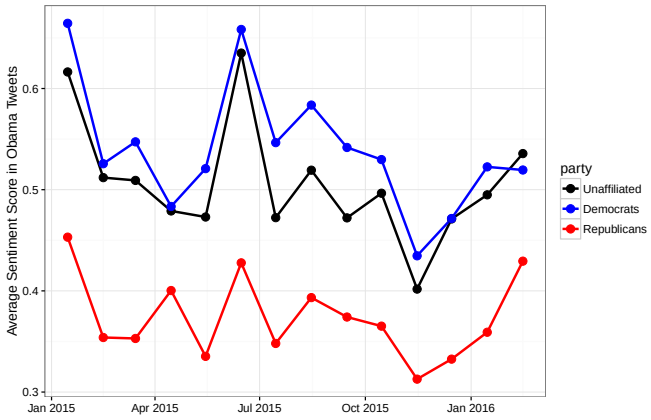


could have a negative tone. Hillary Clinton is the most mentioned Democratic candidate, although Bernie Sanders is not far behind. Ted Cruz is the second most mentioned Republican candidate after Donald Trump, although a potential explanation is that the dataset contains many tweets not related to the politician, given that "cruz" means "cross" in Spanish, which could explain why so many Hispanics appear to mention this politician.

However, the most interesting patterns emerge when we examine different demographic groups. Most of these results are consistent with previous studies: for example, men appear to discuss politics more often, and Republicans mention Obama at higher rates than Democrats (Barberá and Rivero 2014; Barberá 2015a). Hillary Clinton appears to be more popular among older Twitter users with higher income, whereas Sanders is more frequently mentioned by young users with lower income levels. African-Americans appear to be one of the most politically active group: almost 40% of them have mentioned Obama, more than any other group. This difference is consistent through all the other candidates. In the Republican field, Hispanics are disproportionately more likely to mention Marco Rubio. Donald Trump appears to be highly mentioned by all demographic groups.

As suggested by the fact that Republicans mention the President more often than Democrats, the total number of tweets about a political actors is not necessarily a good proxy of popularity. To address this concern, I apply sentiment analysis to detect how the general tone of tweets about Obama varies across individuals of different parties and over time. Although supervised learning generally yields better performance in sentiment tasks, for this preliminary analysis I rely on a standard dictionary approach (Hu and Liu 2004) that computes a sentiment score for each individual tweet based on the number of positive minus negative words that it contains. Despite the simplicity of this method, it appears to achieve good accuracy, as Figure 2 demonstrates. Here, I show the average sentiment score for tweets sent by each party group each month between January 2015 and February 2016. As expected, Republicans tweet more negatively about Obama than unaffiliated voters, and Democrats, respectively.

Figure 2: Sentiment score in tweets mentioning Obama, by month and party



As discussed earlier, one possible method to improve the representativeness of samples of Twitter users is to 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), would allow us to compensate for the fact that, for example, Republicans are underrepresented on Twitter. More in detail, this approach would consist 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:

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

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). In particular, I fit a multilevel logistic regression, where the dependent variable is whether each individual tweet is positive or negative, and the independent variables are each of the sociodemographic variables of the user who published it:

$$Pr(y_i = \text{positive}) = logit^{-1}(\alpha_0 + \alpha^{\text{male}}_{j[i]} + \alpha^{\text{race}}_{j[i]} + \alpha^{\text{party}}_{j[i]} + \alpha^{\text{age}}_{j[i]} + \alpha^{\text{income}}_{j[i]})$$

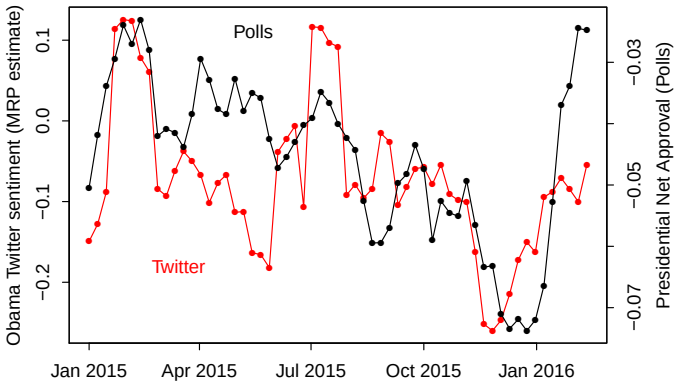
After estimating this model, I compute the predicted proportion of positive tweets in each cell, and then average at the population level using data from the 2012 Congressional Cooperative Election Study (Ansolabehere and Schaffner 2012), which includes all the relevant sociodemographic variables used 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 3 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 approval* for both time series: the total proportion of positive tweets/responses minus the total proportion of negative tweets/responses. 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, there appears to be a somewhat 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 $r = 0.58$

Interestingly, changes on Twitter appear to take place before they're registered in the polls, perhaps suggesting that individuals on Twitter are leading public opinion.

Table 7 provides a more systematic test of the relationship between these these two variables using time series regression. Model 1 shows that contemporaneous values of adjusted Twitter sentiment are correlated with job approval based on surveys. Model 2 shows that this effect is robust

Figure 3: Comparing Twitter- and survey-based measures of presidential job approval



to controlling for previous values of net approval in surveys. Finally, using an error-corrected model Model 3 shows that both lagged values and changes in net approval on Twitter predict current values of net approval according to surveys. All these results are robust to controlling for the total number of tweets per month. While this analysis is still preliminary, it demonstrates that Twitter-based metrics appear to have significant predictive power for a relevant public opinion time series.

Table 7: Time Series Regression

	(1)	(2)	(3)
Net approval (Twitter)	0.15**	0.04**	
	(0.03)	(0.02)	
Net approval (Survey, Lagged)		0.84**	0.86**
		(0.07)	(0.07)
Net approval (Twitter, Diff.)			0.05**
			(0.02)
Net approval (Twitter, Lagged)			0.03*
			(0.02)
Constant	-0.11**	-0.02**	-0.02*
	(0.01)	(0.01)	(0.01)
N	59	58	58
R2	0.35	0.83	0.84

Standard errors in parentheses. Signif.: *10% **5%.
DV: Net presidential approval in surveys.

Inequality in Exposure to Political Information

An important normative and empirical debate in the literature about the political effects of internet use is whether social media usage is creating 'echo chambers'. If individuals are indeed embedded in homogenous networks in social networking platforms, and only exposed to the opinions of like-minded users, it is possible that social media usage could exacerbate political polarization (Sunstein 2001;

Hindman 2008; Huckfeldt 1995; Mutz 2006). Some studies appear to find empirical evidence of strong ideological sorting in Twitter networks. (Conover et al. 2012; Colleoni, Rozza, and Arvidsson 2014). However, when a broader set of individuals and topics are considered, individuals do not appear to be as segregated as previously thought (Barberá et al. 2015). In fact, most social media users are embedded in ideologically heterogeneous networks (Barberá 2015b; Bakshy, Messing, and Adamic 2015).

This set of studies have focused on ideological differences in information consumption and exchange, but Twitter networks could potentially be segregated along other sociodemographic traits. Understanding whether networks are clustered based on these characteristics is relevant because it helps us understand how the use of social media affects social inequalities in information consumption. For example, if young people tend to follow and be followed by other young individuals, more frequent social media usage could lead to significantly less exposure to political information, given that age is correlated with political interest.

To examine this question, I use the panel of Twitter users described in the following section, and quantify what proportion of users were potentially exposed to two popular tweets: Hillary Clinton's presidential run announcement on April 12, 2015, and Ellen Degeneres' selfie tweet at the 2014 Academy Awards. I collected the list of individuals who retweeted each of these tweets in the 24 hours after they were originally sent. This allows me to distinguish whether the individuals in my panel were either *directly* exposed to the tweet (because they follow Hillary Clinton or Ellen Degeneres), *indirectly* exposed to the tweet (because they follow someone who retweeted that tweet), or not exposed.

Table 8 reports the proportion of users that fall in each of the two first categories for both tweets. The results here illustrate the importance of 'inadvertent' exposure (Brundidge 2010): for over 90% of users who were potentially exposed to Clinton's tweet, the tweet appeared in their timeline as a retweeted by one of the other other users they follow. The second broad pattern is that, as expected, there are strong selection effects in direct exposure to this political message: women are 33% more likely than men to follow Hillary Clinton; Democrats are around three times more likely to follow Hillary Clinton than Republicans and affiliated; and adults over 40 years old are also much more likely to follow her than younger Twitter users. However, the crucial lesson from this analysis is that these relative differences are much lower in magnitude once we consider indirect exposure. A third of almost all sociodemographic groups were exposed to this tweet. One exception to this trend is the group of African-American Twitter users: despite being directly exposed to this tweet at similar rates to Hispanics, their rate of indirect exposure is only around 25% vs 37% for Hispanics and 34% for Whites.

The comparison with the tweet by Ellen Degeneres allows me to disentangle whether these patterns are particular to particular messages or more general. As in the previous case, I find that retweets are a powerful mechanism to reduce inequalities in exposure: women are more than three times more likely to follow Ellen, but the rates of total exposure by men and women are relatively similar. While the comparison is difficult given that African-Americans were directly exposed to this tweet at lower rates than Hispanics or Whites, the result for this racial group also suggests the existence of strong clustering along race and ethnicity. Further research is needed, but this finding raises the question of whether social media could actually exacerbate racial inequalities in access to political information.

Table 8: Proportion of users who were directly and inadvertently exposed to Hillary Clinton's presidential run announcement tweet and Ellen Degeneres' selfie tweet during the 2014 Academy Awards.

	Clinton tweet		Ellen 'selfie'	
	Direct	Inadv.	Direct	Inadv.
All	2.1	31.8	10.5	23.6
Men	1.8	29.2	4.8	21.9
Women	2.4	34.3	15.6	25.1
Democrats	3.0	33.2	10.1	22.3
Republicans	1.3	26.5	11.9	21.9
Unaffiliated	0.8	31.9	10.4	27.1
Age 18-25	0.7	26.5	9.7	24.1
Age 26-40	2.6	35.0	11.7	23.4
Age >40	3.3	33.9	9.7	23.3
Low income	1.5	26.9	10.3	20.3
Middle income	1.2	33.7	10.3	25.1
High income	3.8	37.2	10.9	27.0
African-Am.	1.1	25.5	5.3	17.6
Hispanic	1.0	36.6	9.4	27.5
White	4.0	34.2	16.3	26.3

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 examining how the salience of political candidates varies across demographic groups, and how these covariates affect exposure to political messages on Twitter. While further work is needed, these two applications highlight the value of this method to obtain more representative measures of public opinion from Twitter data, and the importance of social media in reducing inequalities in access to political information.

Beyond finishing the data collection, 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 differ-

ent 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, Liu, and Ruths 2012] Al Zamal, F.; Liu, W.; and Ruths, D. 2012. Homophily and latent attribute inference: Inferring latent attributes of twitter users from neighbors. In *ICWSM*.

[Ansolabehere and Schaffner 2012] Ansolabehere, S., and Schaffner, B. 2012. Cces common content, 2012. *Cooperative Congressional Election Study (distributor), version* 2.

[Bakshy, Messing, and Adamic 2015] Bakshy, E.; Messing, S.; and Adamic, L. 2015. Exposure to ideologically diverse news and opinion on facebook. *Science* aaa1160.

[Barberá and Rivero 2014] Barberá, P., and Rivero, G. 2014. Understanding the political representativeness of twitter users. Forthcoming in Social Science Computer Review.

[Barberá et al. 2015] Barberá, P.; Jost, J. T.; Nagler, J.; Tucker, J. A.; and Bonneau, R. 2015. Tweeting from left to right is online political communication more than an echo chamber? *Psychological science*.

[Barberá 2015a] Barberá, P. 2015a. Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data. *Political Analysis*.

[Barberá 2015b] Barberá, P. 2015b. How social media reduces mass political polarization. evidence from germany, spain, and the u.s. In *APSA 2015 Anual Meeting Paper*.

[Beauchamp 2014] Beauchamp, N. 2014. Predicting and interpolating state-level polling using twitter textual dat. In *APSA 2014 Anual Meeting Paper*.

[Blevins and Mullen 2015] Blevins, C., and Mullen, L. 2015. Jane, john ... leslie? a historical method for algorithmic gender prediction. *Digital Humanities Quarterly*.

[Bond et al. 2012] Bond, R.; Fariss, C.; Jones, J.; Kramer, A.; Marlow, C.; Settle, J.; and Fowler, J. 2012. A 61-million-person experiment in social influence and political mobilization. *Nature* 489(7415):295–298.

[Brundidge 2010] Brundidge, J. 2010. Encountering "difference" in the contemporary public sphere: The contribution of the internet to the heterogeneity of political discussion networks. *Journal of Communication* 60(4):680–700.

[Chen et al. 2015] Chen, X.; Wang, Y.; Agichtein, E.; and Wang, F. 2015. A comparative study of demographic attribute inference in twitter. In *Ninth International AAAI Conference on Web and Social Media*.

[Colleoni, Rozza, and Arvidsson 2014] Colleoni, E.; Rozza, A.; and Arvidsson, A. 2014. Echo chamber or public sphere? predicting political orientation and measuring political homophily in twitter using big data. *Journal of Communication* 64(2):317–332.

[Conover et al. 2012] Conover, M. D.; Gonçalves, B.; Flammini, A.; and Menczer, F. 2012. Partisan asymmetries in online political activity. *EPJ Data Science* 1(1):1–19.

[Culotta, Ravi, and Cutler 2015] Culotta, A.; Ravi, N. K.; and Cutler, J. 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 et al. 2016] Diaz, F.; Gamon, M.; Hofman, J. M.; Kıcıman, E.; and Rothschild, D. 2016. Online and social media data as an imperfect continuous panel survey. *PloS one* 11(1):e0145406.

[DiGrazia et al. 2013] DiGrazia, J.; McKelvey, K.; Bollen, J.; and Rojas, F. 2013. More tweets, more votes: Social media as a quantitative indicator of political behavior.

[Florini 2013] Florini, S. 2013. Tweets, tweeps, and signifyin': Communication and cultural performance on "black twitter". *Television & New Media* 1527476413480247.

[Gayo Avello, Metaxas, and Mustafaraj 2011] Gayo Avello, D.; Metaxas, P. T.; and Mustafaraj, E. 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.

[Hampton et al. 2014] Hampton, K.; Rainie, L.; Lu, W.; Dwyer, M.; Shin, I.; and Purcell, K. 2014. Social media and the 'spiral of silence'. *Pew Research Center, Washington, DC pewinternet. org/2014/08/26/social-mediaand-the-spiral-of-silence*.

[Hindman 2008] Hindman, M. 2008. *The myth of digital democracy*. Princeton University Press.

[Hu and Liu 2004] Hu, M., and Liu, B. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, 168–177. ACM.

[Huckfeldt 1995] Huckfeldt, R. R. 1995. *Citizens, politics and social communication: Information and influence in an election campaign*. Cambridge University Press.

[Jungherr et al. 2016] Jungherr, A.; Schoen, H.; Posegga, O.; and Jürgens, P. 2016. Digital trace data in the study of public opinion: An indicator of attention toward politics rather than political support. *Social Science Computer Review*.

[Jungherr, Jürgens, and Schoen 2012] Jungherr, A.; Jürgens, P.; and Schoen, H. 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.

[Jungherr 2015] Jungherr, A. 2015. *Analyzing Political Communication with Digital Trace Data*. Springer.

[Kwak, Moon, and Lee 2012] Kwak, H.; Moon, S. B.; and Lee, W. 2012. More of a receiver than a giver: Why do people unfollow in twitter? In *ICWSM*.

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

[Lin et al. 2013] Lin, Y.-R.; Margolin, D.; Keegan, B.; and Lazer, D. 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*, 737–748. International World Wide Web Conferences Steering Committee.

[Little 1993] Little, R. J. 1993. Post-stratification: a modeler's perspective. *Journal of the American Statistical Association* 88(423):1001–1012.

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

[Mejova, Weber, and Macy 2015] Mejova, Y.; Weber, I.; and Macy, M. W. 2015. *Twitter: A Digital Socioscope*. Cambridge University Press.

[Mislove et al. 2011] Mislove, A.; Lehmann, S.; Ahn, Y.-Y.; Onnela, J.-P.; and Rosenquist, J. N. 2011. Understanding the demographics of twitter users. *ICWSM* 11:5th.

[Mutz 2006] Mutz, D. C. 2006. *Hearing the other side: Deliberative versus participatory democracy*. Cambridge University Press.

[Newman et al. 2008] Newman, M. L.; Groom, C. J.; Handelman, L. D.; and Pennebaker, J. W. 2008. Gender differences in language use: An analysis of 14,000 text samples. *Discourse Processes* 45(3):211–236.

[O'Connor et al. 2010] O'Connor, B.; Balasubramanyan, R.; Routledge, B. R.; and Smith, N. A. 2010. From tweets to polls: Linking text sentiment to public opinion time series. *ICWSM* 11(122-129):1–2.

[Park, Gelman, and Bafumi 2004] Park, D. K.; Gelman, A.; and Bafumi, J. 2004. Bayesian multilevel estimation with poststratification: state-level estimates from national polls. *Political Analysis* 12(4):375–385.

[Pennacchiotti and Popescu 2011] Pennacchiotti, M., and Popescu, A.-M. 2011. A machine learning approach to twitter user classification. *ICWSM* 11:281–288.

[Rao et al. 2010] Rao, D.; Yarowsky, D.; Shreevats, A.; and Gupta, M. 2010. Classifying latent user attributes in twitter. In *Proceedings of the 2nd international workshop on Search and mining user-generated contents*, 37–44. ACM.

[Schober et al. 2016] Schober, M. F.; Pasek, J.; Guggenheim, L.; Lampe, C.; and Conrad, F. G. 2016. Research synthesis social media analyses for social measurement. *Public Opinion Quarterly* nfv048.

[Schwartz et al. 2013] Schwartz, H. A.; Eichstaedt, J. C.; Kern, M. L.; Dziurzynski, L.; Ramones, S. M.; Agrawal, M.; Shah, A.; Kosinski, M.; Stillwell, D.; Seligman, M. E.; et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one* 8(9):e73791.

[Sunstein 2001] Sunstein, C. 2001. *Republic.com*. Princeton University Press.

[Sylwester and Purver 2015] Sylwester, K., and Purver, M. 2015. Twitter language use reflects psychological differences between democrats and republicans. *PloS one* 10(9):e0137422.

[Tumasjan et al. 2010] Tumasjan, A.; Sprenger, T. O.; Sandner, P. G.; and Welpe, I. M. 2010. Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM* 10:178–185.

[Vaccari et al. 2013] Vaccari, C.; Valeriani, A.; Barberá, P.; Bonneau, R.; Jost, J. T.; Nagler, J.; and Tucker, J. 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 et al. 2014] Wang, W.; Rothschild, D.; Goel, S.; and Gelman, A. 2014. Forecasting elections with non-representative polls. *International Journal of Forecasting*.

[Zou and Hastie 2005] Zou, H., and Hastie, T. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67(2):301–320.