Seattle Needles

December 8, 2020

Final Report

To What Extent Can Social Media Predict the 2020 US Presidential Election?

Section 1: Introduction

1.1: Introduction to the Problem

As group 2B, our aim is to study what the social universe can reveal about the real world. We want to know how reliable the social universe is for listening, and how we can calibrate these voices in order to exploit them and learn something about the real world. We have chosen one corner of the social universe: Twitter, and are listening to political voices in order to learn about real political opinion in the United States. In this report, we seek to answer the question: to what extent can social media predict the 2020 United States presidential election?[1]

Our motivation to study this problem is multi-faceted. First, as stated above, we hope that studying this subset of the social universe will give us a better understanding of how we can calibrate voices of the social universe in general. Second, listening to and analyzing political voices on Twitter can help us understand how democracy works in practice. Political scientists, philosophers, sociologists, psychologists, and advertising firms and departments may all be interested in how the social and political worlds interact on Twitter. Results from this study could also affect how politicians conduct their campaigns. Perhaps this study will reveal that politicians need not worry about dissent on Twitter, or conversely it may make them realize how important it is to maintain a positive dialogue on social media. Third, we think that this project will appeal to people who are able to vote in the United States and use social media. We hope that anyone who encounters this project will reflect on how they share their political views on

[1] Our results with interactive visualizations are hosted on https://election2020.web.illinois.edu/.

social media, and analyze how their behaviors align with our results. Lastly, this question in social media analysis and data mining is still unsolved. We hope that we can contribute to solving it by combining state-of-the-art techniques from well known papers with our own ideas.

1.2: Review of Related Work

The goal of this section is to give readers an understanding of how our project relates to and advances on the social media election prediction literature. Additionally, we will discuss related areas used in the literature, such as sentiment analysis, political profiling, demographic profiling, and geolocation.

1.2.a: What Has Been Done Well

Initially, the methods for predicting elections using social media were not very sophisticated. For instance, Tumasjan et al. 2010 just counted the tweets with mentioning the political candidates and assumed the more mentions one has, the more votes they will get in the election [41, 45], which Gayo-Avello 2011 showed performs *worse* than a random classifier in other elections [41]. Indeed, after doing our analysis of 2020 and 2016 tweets, we saw Trump mentioned much more than Biden, even though Biden won the election. Another example of a popular methodology that the literature still uses is O'Connor et al (2010), which used the number of positive and negative sentiment tweets about a political candidate. Then, they computed a sentiment score for each candidate based on the aggregate number of positive and negative tweets about each candidate and assumed that the candidate with the better sentiment score would win the election at the national level [41, 50]. Although this method of counting the number of positive and negative tweets remains popular, the literature has been experimenting with new methods for predicting elections using social media in the last few years. Examples include using time-series analysis, novel distance metrics for high-dimensional sentiment analysis, and features from the Twitter network's graph structure.

The use of social media in the last few decades has been helpful in determining people's attitudes with respect to specific topics or events, and sentiment analysis can be really useful as mentioned before. Researchers have begun to think about how to use sentiment analysis to identify the population's attitude on some topics. Sentiment analysis of text can help determine the polarity and inclination of people toward a specific topic, item or entity. A wide research interest in natural language processing studies the hypothesis based on the users' sentiment on social media. Most researchers have built sentiment classifiers and classified the data basically as positive, negative and neutral. A two stage framework can be formed to create a training data from the mined data and to propose a scalable machine learning model to predict the election result. There are two main kinds of sentiment classifiers, a Machine Learning based approach and Lexicon based approach. Both of these approaches have been used in many other fields of application. The ML based approach uses popular text classification algorithms like Naive Bayes and SVM, which are Supervised Learning Algorithms that require labeled training data. In order to train the model, we need to numerically represent the preprocessed data. The well-known techniques for vectorization of words in Natural Language Processing are CountVectorization and Tf-IDF transformation. The Lexical approach uses sentiment dictionaries with opinion words and matches them with data to determine polarity. Sentiment values are manually assigned to words that describe the positive, negative and neutral attitude of the speaker. The total polarity of a tweet is then calculated by adding the scores of all the individual words. These counts are then used to determine the percentage of positive, negative and neutral tweets. The polarity will be normalized in ranges between [-1,+1]. To select the appropriate approach before the project starts, we compared these two approaches. Machine learning methods, such as SVM and Naive Bayes have the highest accuracy and can be regarded as the baseline learning methods, while lexicon-based methods are very effective in some cases, which require little effort in human-labeled documents. But ML based approach needs sufficient and particular training data. In conclusion, when we can't obtain the training data with labels in a specific domain, lexicon-based methods perform better than machine learning methods [34].

The research focus on election prediction using social media in the past few years has begun using these two approaches in different ways. Tweets are classified as positive or negative, and people are also trying to classify tweets into many different emotional categories. However, most studies use positive and negative classification to make predictions. Some studies have done sentiment analysis and categorization of tweets for the 2016 election, and statistically summarized the results. If classifying tweets as positive or negative, Clinton would win the election since she has more positive tweets than Trump [37]. If classifying tweets into 6 different sentiment categories, Trump will get benefit of joyous tweets and Trump has more tweets on all other emotions. A large part of Clinton tweets express sadness [38], but we still can't make a prediction directly from the result. Thus, it is difficult for us to make accurate predictions about the election simply by counting.

As we will discuss further below, analysis of demographic characteristics of social media users is generally missing from the election prediction literature. However, much work has been done on classification of users outside of the election prediction topic. We were interested in classifying users by location, gender, and race/ethnicity, so we searched the literature for these topics.

We found three common methods for locating users: 1) using the self-reported location of users [6,10], 2) using the location of tweets determined by GPS, cell towers, or manual selection [7], and 3) using the self-reported location of users' friends [8,9]. Each of these methods have pros and cons which we will discuss briefly. Method 1 is simple, and according to Mislove in [6], 75% of users are labeled. However, the accuracy of these locations is unknown, and specificity of location varies between users. Method 2 is more accurate than 1 since it involves measurement of location, but only 1% of tweets are geolocated [7]. We should also keep in mind that the location of tweets may not imply "home" location of the user. Method 3 is less prone to being skewed by inaccurate locations than 1 because it involves using multiple self-reported locations to infer another. Another pro to this method is that ~50% of users can be located using their friends' locations alone [8], but the collection of this network data is costly.

To classify users by gender, we found the most straightforward way was detailed in [6,11], which is looking at first names of users. Other less common methods include color-based classification [12] and text and image processing [13]. Profiling users by race was also done in [6] by last name, and this was the most robust method we could find.

On another note, we will briefly discuss how the literature has become self-conscious of its own methodological mistakes and is improving. For instance, Gayo-Avello (2013) made many criticisms about the literature's conceptualization of Twitter data, making four main points:

1. Twitter users who tweet about politics are self-selected from the entire population, which overlaps with the sample of the entire population who voted in the election because not everyone in the population votes or tweets about politics.

2. Certain demographic and political preference groups are over/under-represented on Twitter depending on the location. Therefore, there is a need to over/under-sample and weigh these groups differently.
3. Not every account that tweets about politics is a potential voter because they may be a bot account that spams about politics.
4. Some human accounts are not potential voters because they are not of voting age, not politically involved, or a troll account.

Most papers have not implemented or considered any of these methodological issues mentioned by Gayo-Avello (2013). However, our project addresses (1) and (2) by reconceptualizing geolocated political tweets as a voluntary, self-selected web survey with exit polls as a reference sample and following the weighting scheme in Bethlehem (2010), and some parts of (3) by running a bot detector on a random sample of our tweets. Moreover, a recent paper, Yang et al. (2020), addresses (4) by considering potential voters who used the #ivoted hashtag or tweeted photos related to politics.

1.2.b: What is Lacking

So far, the social media election prediction literature has only used the most simple methods for Twitter data collection. Some papers use a predefined list of search terms related to politics about politics, such as obama, romney, pelosi, reid, biden, mcconnell, cantor, boehner, liberal, liberals, conservative, conservatives, republican, republicans, democrat, and so on [42]. However, this approach would fail if there are new and emerging keywords related to political trends. A few papers find keywords and hashtags that are highly correlated with political polls, then filter out non-political ones [43]. However, the election polls may be unreliable, and this approach still necessitates the ability to distinguish between political and non-political terms manually. We propose using online trend analysis with topic models, which uses an online Latent Dirichlet allocation (LDA) system [44] as the best way to achieve this goal. However, this has never been implemented in any election prediction paper yet. In general, little work has been done to find the optimal parameters for querying beyond search terms. For instance, no paper has ever analyzed the optimal time to collect tweets before an election, if there is one. Nevertheless, the median length of this data collection window is about 8 weeks, ranging from 1 week to 8 months.

Next, regarding the location of tweets, no paper has analyzed what the best way of getting tweets about a certain location from potential voters. Some papers, including this one, filter tweets by their geocode or the location specified on a Twitter profile (used by Gayo-Avello 2011 and Skoric et al. 2012). To augment the specified Twitter users and tweets with specified locations, we could also augment this dataset by using a data mining approach to geolocate users based on their following relationships. However, to our knowledge, this has never been implemented in the literature. On a related note, no paper has analyzed the difference of the subset of political tweets that are geolocated versus the full set of tweets. Lastly, some others (Tumasjan et al. 2010; Jungherr et al. 2011; Tjong Kim Sang & Bos 2012) make an assumption about a user's location based on the language(s) used in their tweets [45, 46, 47, 48, 49]. However, this is not possible for elections that are on the world stage discussed widely by many countries, such as the 2020 United States presidential election.

Only using the simplest of all sentiment analysis methods (positive / negative) may not be enough to analyze the attitude of a vast population. Some other research using network structure to predict people's attitude has a higher accuracy [57]. The result shows that techniques based on the statistical analysis of political communication networks provide higher accuracy than text-content based approaches, but this method needs human labeled users. At the same time, we can see from the method of using 6 different emotions to predict the 2016 general election [38] that using more emotion classifications can more reasonably analyze the final results. More research still needs to be done to understand how to use this method to achieve a more accurate prediction. In addition, because of the diversity of languages, sentiment analysis is often misled. The way people express emotions is too complicated. It is not possible to judge the emotions expressed by people simply by text. Some studies have shown that if people's facial expressions in pictures are used to judge emotions at the same time, higher accuracy can be obtained [39].

1.2.c: How Our Contributions Advance the Literature

The approach for data collection in this project advances on the literature for using social media to predict political elections in the United States in three significant ways. First, this project queries tweets from each state, enabling the prediction of the electoral college. Second, this project uses a *user-centric* paradigm, rather than a *tweet-centric* one to address the four main methodological critiques made by Gayo-Avello (2013) that were mentioned in section *1.2.a*. Third, this project approximates the sampling bias of political preferences due to self-selection. In other words, the only people sampled were those who voluntarily tweeted about a political candidate, rather than being randomly sampled from the entire population of voters of a certain demographic group within a particular state.

Section 2: Approach

2.1: Data Collection

In this section, we will discuss how we collected data from Twitter and exit polls about the 2016 and 2020 election for each state and demographic group.

2.1.a: Twitter Data

Our approach can be briefly summarized as follows. We collected geolocated tweets about the Democratic and Republican presidential candidates from the 2016 and 2020 presidential elections that were collected from each city and state in the United States within a 13 week window ending the day before the date of the election using the Twitter API. Specifically, regarding the 2020 election tweets, a query was made for all tweets containing “biden” or “trump” geolocated at a particular city in each state and the state itself from Tuesday, August 3, 2020 at 12:00:01 pm Eastern Standard Time until Monday November 2, 2020 at 11:59:59 p.m. Eastern Standard Time. In the end, we obtained 247,837 unique tweets that were geolocated and contained at least one of the query terms. Of those tweets, the proportion of each political preference was 28% neutral, 36% Democrat, 34% Republican, 2% unknown. To collect data from the 2016 election, we followed an analogous procedure. A query was made for all tweets containing “trump” or “hillary” or “clinton” geolocated at a particular city in each state and the state itself from Tuesday, August 9, 2016 at 12:00:01 pm Eastern Standard Time until Monday November 7, 2016 at 11:59:59 p.m. Eastern Standard Time. The end result was only 67,867 geolocated tweets with at least one of the candidates' names, which is about 25% of the number of tweets collected from 2020. Of those 67,876 tweets, 40% were neutral, 29% Democrat, 26% Republican, and 5% unknown. One thing to note is that the geolocated tweets containing one of the candidate’s names were more politically polarized in 2020 than in 2016, based on the percent of neutral tweets about a candidate. Interestingly, we will see later on in the Results and Analysis sections of this paper how different distributions of political tweets (i.e. neutral, Democrat, Republican, unknown) of a particular demographic group or state about a political party have a detrimental effect on the prediction results for that demographic group or state. Next, we will briefly discuss some justifications for three design decisions we made related to (1) our queries, (2) our decision to use geolocated tweets, and (3) our method for profiling the politics of users.

First, limiting the queries to the names of the political candidates themselves was a pragmatic decision necessitated by the time limits imposed by the semester course project structure to ensure there was enough time for comprehensive analysis of the data. In particular, even after simplifying our queries, the data collection query took more than one week to complete. Several additional steps could have been implemented for a more comprehensive query. First, a more comprehensive set of queries could have included additional political terms such as “mcconnel”, “conservative”, “republican”, and so on. Second, Google Trends could have been used to find related terms to known terms, as well as political terms highly correlated with poll numbers. Third, an online trend analysis topic model system using Latent Dirichlet Allocation (LDA) could have been used to find emerging political keywords and hashtags.

Second, our decision to use geolocated tweets significantly reduced the number of tweets collected, as geolocated tweets are only about 1% of all tweets [7]. Nevertheless, we believe this was necessary because hashtags and news stories related to Donald Trump were often top trending topics on Twitter during the election. Moreover, this allowed us to advance on the implicit assumption made in many papers in the literature that tweets about political candidates are from a particular country. We could have supplemented our data by using geolocation methods based on the user’s follower-followee relationships.

Third, we will justify our decision to profile the political alignment of users based on using sentiment analysis to determine if they had a positive or negative sentiment towards the Democrat or Republican presidential candidate. This was another pragmatic decision to ensure that we could finish our analysis in time, because the alternative would have been a massive undertaking. In particular, the alternative would have been to use the follower-followee relationship to determine the political alignment of every Twitter user in the United States (either geolocated with a data mining method, specified in their Twitter profile, or with geolocated tweets turned on). This would have significantly increased the amount of queries that we would have needed to make because we would have had to consider every USA-based Twitter user's follower-followee relationships to profile their politics, not just those who tweeted about politics.

2.1.b: Exit Poll Data

Exit polls are used for three main reasons. First, voting is anonymous, so we have no other way to approximate what proportion of each demographic group in each state voted for a certain presidential candidate. Second, surveys done prior to the election are subject to changes in the political preferences of the electorate until the election. Third, surveys conducted prior to the election are often smaller in scale and thus have a larger margin of error. Therefore, exit polls are the best source of information for estimating what percent of voters of a certain demographic group that live in a certain state voted for a certain presidential candidate. Later, we will discuss how we use exit polls to estimate the sampling bias of the political preference of a certain demographic group in a state due to self-selection of political tweeters.

The exit poll data was scraped from CNN's 2016 and 2020 exit polls.[2] We used CNN's exit polls because they are the most comprehensive with respect to race / ethnicity, gender, and both race / ethnicity and gender (i.e. white, women, and white women, respectively). Moreover, CNN's exit polls survey the entire nation and 25 states.

However, CNN's exit poll data is incomplete in various ways. First, the exit polls data only has full data about gender breakdowns for 20 states, full data about race / ethnicity

[2] The exit polls for the national and about 25 states can be found here:
1. https://www.cnn.com/election/2016/results/exit-polls/national/president
2. https://www.cnn.com/election/2020/exit-polls/president/national-results.

breakdowns for 3 states, and full data about both race / ethnicity and gender for 1 state, but data about white people, white men, and white women were available for all states that were surveyed. As a result, this posed a major methodological problem for our approach because we were not able to use exit polls for the demographic groups that were not surveyed in the exit polls. Moreover, the data was very incomplete for Hispanic / Latino people, black people, asian people, Hispanic / Latino men & women, black men & women, and asian men & women, which reduced our ability to predict their respective voter shares for each party in each state. In the Results and Analysis sections, it will be highlighted how these groups performed worse than their white, men, white men, and white women counterparts in many cases.

2.2: Sentiment Analysis

After collecting tweets, using the sentiment analysis discussed before, we classified the tweets into pro-Democrat or pro-Republican. At the beginning, as the ML based sentiment analysis has more accurate prediction, we trained a model based on the 2016 election using the SVM algorithm. But when we did a manual test we found that the predictions for tweets about Biden are prone to drift from the correct sentiment which is caused by the inconsistency error from ML based method. Also since the ML based method took more time to predict, we decided to change our approach to Lexical based method which is more general and not context-dependent. The Lexical based method is also much faster than the ML based approach. The classifier contains three main steps. In the first step, we performed an analysis to define which tweets are about politics and which tweets are not about politics. In the second step of the proposed methodology, we performed the tweets’ texts sentiment analysis which has two distinct approaches: one for political tweets that were about both candidates at the same time and another one for tweets that were about only one candidate. The political users’ timelines and sentiment analysis allowed us to classify each user. Thus, in the third step of the methodology, we arranged users into three different classes: *Trump Supporter, Biden Supporter,* and *Neutral*.

First step: Political analysis. In order to know whether the content of tweets is politically relevant or not, we use three checks naively. The tweets not related to politics will be automatically filtered out.

1. The tweet targets at least one candidate or party; (Trump, #Trump)
2. The tweet mentions at least one candidate or party; (@realDonaldTrump)
3. The tweet has a candidate's proper name or party; (DT, Rep)

Second step: Pre-sentiment analysis. After political tweet identification, we observed in the data set that political tweets involving two candidates at the same time are a challenge to analyze their political sentiment. Because the full text sentiment analysis only considers the sentiment of users when they post a tweet, not the user's opinion on each topic in the tweet. To solve this problem, we divide sentiment analysis into two parts. In the first part, we considered the entire tweet and performed sentiment analysis. Therefore, the text sentiment is assigned to the candidate if and only if the tweet is about only one candidate. In the second method, we determine the words related to each candidate if and only if the tweet involves two candidates. Therefore, we can only consider words related to them for sentiment analysis. In order to extract the relevant words, we use Stanford NLP toolkit to get the parse tree so that we can identify the words related to the candidates.

Third step: User political preference classification. After we know which part of the tweet should be used to do sentiment analysis, we can determine the political preference based on it. There are two situations. If the tweet mentions only one candidate, the sentiment is assigned based on the whole text. If the candidate is from the Democratic Party, when the sentiment is positive, we judge that this tweet is in favor of the Democratic Party. In another case, when the tweet mentions both of the candidates, we assign sentiment scores to both candidates based on their relevant words. If both are positive or negative or both are neutral, we assign it a neutral state. If not, we just compare the score and use the larger one to determine the political polarity.

2.3: User Profiling

After collecting tweets (along with information about users including screen name and Twitter handle) using the methods described above, we wanted to organize the users into groups that would aid in our listening to the social universe. Since Twitter does not provide data on its users' demographics, we had to employ other methods from the literature to label our users. We chose to label the users with a gender and race/ethnicity as done by Mislove (2011) and described in section *1.2. Review of Related Work*. These methods were simplest for us and did not require any additional data collection from Twitter (network information, a large body of tweets, etc.) like some other methods did. We followed Mislove's methods exactly, beginning by splitting users' screen names at a space character into assumed first and last names. We labeled users with a gender (male or female) by consulting the top U.S. baby names since 1880, and with a race/ethnicity (White, Black, Asian/Pacific Islander, American Indian/Alaska Native, or Hispanic) by consulting the top last names in the 2010 U.S. census. [3]

2.4: Bot Detection

We use a tool online called Botometer. It is a random forest based algorithm that will help determine the likelihood of a user account being a bot account. Since the tool sets a quota limit for access, we randomly sampled 434 users (half Democrat, half Republican) from our

---

[3] The raw U.S. baby name and census data can be found here:
1. https://www.kaggle.com/kaggle/us-baby-names?select=NationalNames.csv
2.https://www.census.gov/topics/population/genealogy/data/2010_surnames.html

database which are responsible for around 1,000 tweets. After running the algorithm, only 13 out of those 434 sampled users (almost 3%) were identified as more than 50% likely to be bots, with the bots having a similar distribution of political alignment to the distribution of Democrat and Republican Twitter users. Although there are more comprehensive steps we could take to analyze the effect of bots on our Twitter data, we conclude from this random sample of 434 users from our userbase that that bots are not likely to make an impact on our analysis and prediction results because bots are almost equally represented for each political party.

2.5: Answering the Research Question

To answer our research question, "To what extent can social media predict the 2020 presidential election?", we focus on analyzing the sampling biases of different demographic groups in each state, rather than building the best possible predictor. To get to the heart of our research question, we conceptualized geolocated tweets about a certain presidential candidate on Twitter as a voluntary web survey and used exit polls as a reference sample to calculate the weights for mitigating sampling bias due to self-selection with respect to political preferences. Moreover, by calculating the mean average error (MAE) of each demographic group across all states (explained below step-by-step), we were able to approximate the biases of estimating the true voter share for a political party of each demographic group surveyed in the United States.

The following procedure summarizes our steps to the approximate the sampling bias:

1. Collect the exit polls for 2016 and 2020, which directly gives the estimated voter shares for 2016 and 2020: $voterShareExitPoll2016_{ijk}$ and $voterShareExitPoll2020_{ijk}$, where i is the demographic group, j is the state, and k is the political party.
2. For each state, collect the geolocated tweets containing the search term from that state for both 2016 and 2020.

3. For each user, profile their political preference and demographic group for both 2016 and 2020.
4. Calculate the Democrat and Republican voter shares from 2016 and 2020 for each demographic group in each state: $voterShareTwitter2016_{ijk}$ and $voterShareTwitter2020_{ijk}$ where *i* is the state, *j* is the demographic group, and *k* is the political party. This is simply the percent of each demographic group in each state with a Democrat or Republican political preference.
5. Calculate the weight $weight_{ijk} = voterShareExitPoll2016_{ijk} / voterShareTwitter2016_{ijk}$ for the 2020 Twitter data using the 2016 exit polls and the 2016 Twitter data, where *i* is the state, *j* is the demographic group, and *k* is the political party.
6. Calculate the predicted voter share by multiplying the raw 2020 Twitter voter share by the weight: $voterShareEstimated2020_{ijk} = voterShareTwitter2020_{ijk} * weight_{ijk}$, where *i* is the state, *j* is the demographic group, and *k* is the political party.
7. Calculate the absolute error of the voter share using the 2020 exit polls and 2020 tweets: $voterShareAbsError_{ijk} = |voterShareEstimated2020_{ijk} - voterShareExitPoll2020_{ijk}|$ where *i* is the state, *j* is the demographic group, and *k* is the political party.
8. Calculate the mean absolute error of the voter share for demographic group *j* for party *k*: $voterShareMeanAbsError_{jk} = \frac{1}{n} \sum_{i=1}^{n} voterShareAbsError_{ijk}$ where *i* is the state.
9. Use the $voterShareEstimated2020_{ijk}$ using estimates by political scientists for the proportion of voters in state *i* that are of demographic group *j* for party *k* and the respective $shareOfElectorate_{ijk}$ to predict the outcome of the election using the formula:

$$winnerEstimated_i = argmax_k(percentOfVotes_{i,k = Democrat}\ ,\ percentOfVotes_{i,k = Republican})$$

where

$$percentOfVotes_{i,k} = \sum_{j=1}^{n} voterShareEstimated2020_{ijk} * shareOfElectorate2020_{ijk}\ .$$

10. Determine which elections were predicted correctly and incorrectly using the state popular vote, electoral college, national popular vote across swing states, non-swing states, red states, and blue states.

Section 3: Results

3.1: Results from Demographic Group Resampling

Table 1 shows the uncorrected vote (Republican or Democrat) of each state after following the steps detailed in the approach above, as well as votes corrected by various metrics. The “Gender-Corrected Vote” column contains each state’s vote after men’s and women’s votes were weighted according to what percent of the electorate men and women were in 2016 (45 and 55 % respectively). The same was done for racial/ethnic groups with 74% of the electorate being White, 10% Black, 10% Hispanic, and 5% other or mixed race. Then these two metrics were combined to correct for race and gender, with 33% of the electorate being White men, 41% White women, 4% Black men, 6% Black women, 5% Hispanic men, 5% Hispanic men, and 5% all other groups.[4] Green cells in the table represent a correct prediction, while red cells are incorrect. The last rows show the number of states predicted correctly for each correction method, and how many electoral votes were given to democrats according to the prediction. Note: 6 cells were predicted to be neutral. We replaced these 6 cells with the next most likely vote (Republican or Democrat).

[4] Found here:
https://www.pewresearch.org/politics/2018/08/09/an-examination-of-the-2016-electorate-based-on-validated-voters/

Table 1

State	Uncorrected Vote	Gender-Corrected Vote	Race-Corrected Vote	Race and Gender-Corrected Vote	True Vote	Electoral Votes
Alabama	R	R	R	R	R	9
Alaska	D	D	R	R	R	3
Arizona	D	D	R	D	D	11
Arkansas	D	D	D	D	R	6
California	D	D	D	D	D	55
Colorado	D	D	R	D	D	9
Connecticut	D	D	R	D	D	7
District of Columbia	D	D	D	D	D	3
Delaware	D	D	D	D	D	3
Florida	D	D	D	D	R	29
Georgia	D	D	D	D	D	16
Hawaii	R	D	D	R	D	4
Idaho	D	R	R	R	R	4
Illinois	D	D	D	D	D	20
Indiana	D	D	D	D	R	11
Iowa	D	D	D	D	R	6
Kansas	D	D	D	D	R	6
Kentucky	D	D	R	R	R	8
Louisiana	D	D	D	D	R	8
Maine	D	D	D	D	D	4
Maryland	D	D	D	D	D	10
Massachusetts	D	D	D	D	D	11
Michigan	D	D	D	D	D	16
Minnesota	D	D	R	D	D	10
Mississippi	D	D	D	R	R	6
Missouri	D	D	D	D	R	10

Montana	D	R	D	D	R	3
Nebraska	R	D	D	D	R	5
Nevada	R	R	R	D	D	6
New Hampshire	R	R	R	R	D	4
New Jersey	D	D	D	D	D	14
New Mexico	D	D	D	R	D	5
New York	D	D	D	D	D	29
North Carolina	D	R	R	R	R	15
North Dakota	D	D	D	D	R	3
Ohio	D	D	D	R	R	18
Oklahoma	D	D	D	D	R	7
Oregon	D	D	D	D	D	7
Pennsylvania	D	D	D	D	D	20
Rhode Island	D	D	D	D	D	4
South Carolina	D	D	R	D	R	9
South Dakota	D	D	D	D	R	3
Tennessee	R	R	R	R	R	11
Texas	D	D	D	R	R	38
Utah	D	D	D	D	R	6
Vermont	D	D	D	D	D	3
Virginia	D	D	D	D	D	13
Washington	D	D	D	D	D	12
West Virginia	R	R	R	R	R	5
Wisconsin	D	D	R	R	D	10
Wyoming	D	D	D	R	R	3
Count Correct	27	30	27	33	50	
Dem Electoral Votes	494	481	417	395	306	
National Prediction	D	D	D	D	D	

3.2: Political Bias Correction

We found that correcting by race/ethnicity and gender was not sufficient for prediction results, so we also corrected the political bias of Twitter users due to self-selection. Highlighted rows contain at least one cell that is within the margin of error (~5%) for traditional political surveys. As noted previously, keep in mind that most states had exit poll data for white people, white men, and white women, whereas most did not have data for Hispanic / Latino people, black people, asian people, Hispanic / Latino men & women, black men & women, and asian men & women,

Table 2

State	Predicted Men Percent Democrat in 2020	Predicted Men Percent Republican 2020	Actual Men Percent Democrat 2020	Actual Men Percent Republican 2020	Men Percent Democrat Absolute Error	Men Percent Republican Absolute Error
National	39.55804	54.75814	47	51	7.441964	3.758143
Arizona	39.13218	65.26756	48	50	8.867816	15.26756
Colorado	36.03428	49.85234	49	47	12.96572	2.852342
Florida	43.46951	56.3023	45	54	1.530486	2.302295
Georgia	32.82603	67.1462	43	55	10.17397	12.1462
Iowa	42.5634	100	39	58	3.5634	42
Kentucky	18.96468	81.11212	36	60	17.03532	21.11212
Maine	50.66702	86.18133	41	51	9.667016	35.18133
Michigan	45.28214	52.39443	43	54	2.282141	1.60557
Minnesota	53.53329	58.91162	47	50	6.53329	8.911617
Nevada	82.97305	100	46	51	36.97305	49

New Hampshire	34.6533	65.82891	47	52	12.3467	13.82891
North Carolina	38.98174	56.35886	45	54	6.018263	2.358865
Ohio	34.83211	56.2672	41	57	6.167889	0.732796
Oregon	33.34425	39.3555	51	45	17.65575	5.644502
Pennsylvania	34.9854	60.87848	44	55	9.014604	5.87848
South Carolina	39.79644	72.29128	41	57	1.203562	15.29128
Texas	29.63594	53.73046	40	57	10.36406	3.269542
Virginia	37.01522	55.49547	49	48	11.98478	7.495469
Washington	37.03672	45.02625	48	46	10.96328	0.973753
Wisconsin	29.68041	46.60749	44	54	14.31959	7.392514

Table 3

State	Predicted Women Percent Democrat 2020	Predicted Women Percent Republican 2020	Actual Women Percent Democrat 2020	Actual Women Percent Republican 2020	Women Percent Democrat Absolute Error	Women Percent Republican Absolute Error
National	57.95336	44.13172	55	44	2.953363	0.131723
Arizona	50.18888	32.64955	51	48	0.81112	15.35045
Colorado	100	61.40126	61	37	39	24.40126
Florida	49.36755	68.63898	51	48	1.632446	20.63898
Georgia	38.55918	32.37583	54	45	15.44082	12.62417
Iowa	40.0029	36.05793	51	48	10.9971	11.94207
Kentucky	65.38004	73.64815	36	63	29.38004	10.64815
Maine	36.868	25.73933	61	38	24.132	12.26067

Michigan	39.88151	32.42482	57	42	17.11849	9.575183
Minnesota	56.46527	48.60815	58	41	1.534726	7.608152
Nevada	40.73934	41.76301	54	44	13.26066	2.236987
New Hampshire	79.32919	47.9584	58	40	21.32919	7.958403
North Carolina	55.42091	57.45811	53	46	2.420907	11.45811
Ohio	44.75869	49.59869	53	47	8.241306	2.598693
Oregon	65.33199	32.63247	63	35	2.331993	2.367526
Pennsylvania	65.97791	45.70912	55	44	10.97791	1.709119
South Carolina	35.91222	35.48898	45	53	9.087785	17.51102
Texas	60.75877	59.88218	51	48	9.758771	11.88218
Virginia	63.55129	38.28254	61	38	2.551287	0.282539
Washington	100	66.64281	67	33	33	33.64281
Wisconsin	70.52609	52.07521	56	43	14.52609	9.075208

Using the estimates that men make up 45% of the electorate and women make up 55% of the electorate by studies done by political scientists (not from Twitter data), we estimate the popular vote for the elections below and obtain the following prediction results based on the gender corrections. Note that the green highlight means the election was correctly predicted, and the red highlight means the election was incorrectly predicted.

Table 4

Election	Predicted Democrat Popular Vote Percent	Actual Democrat Popular Vote Percent	Democrat Absolute Error	Predicted Republican Popular Vote Percent	Actual Republican Popular Vote Percent	Republican Absolute Error	Predicted Winner	Actual Winner
National	49.67547	51.3	1.6245	48.9136	46.9	2.01361	D	D
Arizona	45.21337	49.4	4.1866	47.3277	49.1	1.77235	R	D
Colorado	71.21543	55.4	15.815	56.2043	41.9	14.3043	D	D
Florida	46.71344	47.9	1.1866	63.0875	51.2	11.8875	R	R
Georgia	35.97926	49.5	13.521	48.0225	49.3	1.27750	R	D
Iowa	41.15512	44.9	3.7449	64.8319	53.1	11.7319	R	R
Kentucky	44.49313	36.2	8.2931	77.0069	62.1	14.9069	R	R
Maine	43.07756	53.1	10.022	52.9382	44	8.93823	R	D
Michigan	42.3118	50.6	8.2882	41.4111	47.8	6.38886	D	D
Minnesota	55.14588	52.4	2.7459	53.2447	45.3	7.94471	D	D
Nevada	59.74451	50.1	9.6445	67.9697	47.7	20.2697	R	D
New Hampshire	59.22504	52.7	6.5250	56.0001	45.4	10.6001	D	D
North Carolina	48.02328	48.6	0.5767	56.9634	49.9	7.06345	R	R
Ohio	40.29173	45.2	4.9083	52.5995	53.3	0.70048	R	R
Oregon	50.93751	56.5	5.5625	35.6578	40.4	4.74217	D	D
Pennsylvania	52.03128	50	2.0313	52.5353	48.8	3.73533	R	D
South Carolina	37.66012	43.4	5.7399	52.0500	55.1	3.04999	R	R
Texas	46.7535	46.5	0.2535	57.1139	52.1	5.01391	R	R
Virginia	51.61006	54.1	2.4899	46.0283	44	2.02836	D	D
Washington	71.66652	58	13.667	56.9153	38.8	18.1154	D	D
Wisconsin	52.14554	49.4	2.7455	49.6147	48.8	0.81473	D	D

We found that the political bias correction with the gender breakdown of the electorate into men and women works best. We suspect this is due to black, hispanic, asian, and “other” race/ethnicity categories have a high mean absolute error of voter share because they are not represented well with respect to exit polls or geolocated political tweets on Twitter in many states, compared ot white men, white women, and white people. Note that this prediction is using the gender breakdown of the electorate into men and women. There was not enough exit poll data to separately break down the by race/ethnicity and gender and race/ethnicity for each state.

3.6: Website

We built a website as the visualization tool of our final predicted result. The website is hosted at election2020.web.illinois.edu. On top left, users can choose different prediction results which correspond to different stages of correction done on our prediction. Also, each state on the map can be clicked and after clicked, a new page will appear to give the details about that state, including voter turnouts, predicted voter demographics by each race/gender.

Section 4: Analysis

Using demographic group corrections, Table 1 shows that the combined race and gender correction was the most accurate, correctly predicting 33/51 states. While this is still a low percent correct (64.7%), it is a marked improvement from the uncorrected prediction, which was correct for 52.9% of states. This correction was also the most accurate at predicting the number of electoral votes for Joe Biden. The correct number was 306, while the predicted number was 395, which is an error of 29.1%. This is a large improvement from the uncorrected prediction, which predicted 494 electoral votes for Biden, an error of 61.4%. Using the political bias corrections, the popular votes of 15/19 states were correctly predicted and the national popular vote was correctly predicted.

Table 5

Demographic Group	Predicted Percent Democrat National	Actual Percent Democrat National	Absolute Error Democrat National	Mean Democrat Absolute Error All States	Share of the Electorate
men	39.55804	47	7.441964	10.33679	45
women	57.95336	55	2.953363	12.88029	55
white	40.89318	41	0.106824	6.891773	74
black	100	87	13	18.11842	10
hispanic	95.53621	65	30.53621	27.12102	10
asian	100	61	39	20.23952	5
whiteMen	33.17367	38	4.826328	9.267431	33
whiteWomen	48.35895	44	4.358954	10.16306	41
blackMen	100	79	21	21	4
blackWomen	69.11765	90	20.88235	50.84118	6
hispanicMen	88.75278	59	29.75278	28.59573	5
hispanicWomen	100	69	31	16.28148	5

Table 6

Demographic Group	Predicted Percent Republican National	Actual Percent Republican National	Absolute Error Republican National	Mean Republican Absolute Error All States	Share of the electorate

men	54.75814	51	3.758143	12.23825	45
women	44.13172	44	0.131723	10.7573	55
white	71.84741	58	13.84741	19.06028	74
black	8.231564	12	3.768436	6.859984	10
hispanic	38.13247	32	6.132474	5.918734	10
asian	26.79089	34	7.209113	5.077314	5
whiteMen	74.65962	61	13.65962	21.37866	33
whiteWomen	57.92323	55	2.923232	12.3022	41
blackMen	14.62045	19	4.379552	4.379552	4
blackWomen	2.823529	9	6.176471	4.945378	6
hispanicMen	41.33031	36	5.330313	4.397026	5
hispanicWomen	43.04295	30	13.04295	10.69657	5

Using the best demographic group correction for each state category, 8/12 swings states were correctly predicted, 9/22 red states were correctly predicted, and 17/17 blue states were correctly predicted.

Using the political bias corrections, 6/11 swing states were correctly predicted, and 8/8 non-swing states were correctly predicted. 7/12 elections that voted for Democrats were correctly predicted, and 7/7 elections that voted for Republicans were correctly predicted. Again, only 19 states and the entire nation had exit poll data available needed to make this political bias correction.

Section 5: Conclusions

The first conclusion we reached after analyzing the results is that uncorrected tweet sentiment used to label users is an insufficient method to predict election results. After looking at the demographic group correction alone, we determined that this method is also not sufficient for election prediction. These results led us to believe that the self selection bias on Twitter is not due only to misrepresentation of certain demographic groups on the platform, but also a clear ideological bias It is clear from Table 1 that predictions are heavily biased toward a Democratic vote. We were able to correctly predict 100% of blue states, but only 40.9% of red states using this method. Therefore, to partially answer our question, raw tweet sentiment and demographic group-corrected sentiment can be used to predict Democratic state results.

By analyzing the mean average error for each demographic group, we see that for predicting the voter share of Democrats: white people, men, women, white men, and white women have the lowest mean average error across all states and the national popular vote. For predicting the voter share of Republicans, black people, hispanic people, asian people, men, women, white women, black men, black women, and hispanic men have the lowest mean average error across all states and national popular vote have the lowest mean average error.

The ability of social media to predict election outcomes in states depends on the demographic group and the state that the user is in. Since not all demographic groups are represented well on Twitter and exit polls, the political bias correction had mixed results depending on the users.

Section 6: Limitations and Next Steps

Our study had several limitations. Regarding data collection, we had to drastically simplify the number of queries because of the prohibitive amount of time it took to complete the queries for all cities in all states. Additionally, we couldn't get county-level results in time because Twitter API approved late, so we didn't have the ability to query county-specific tweets using geo-polygons. Moreover, we never figured out how to handle the fact that different groups could tweet differently in two different elections. For example, in 2016, the tweets were 40% Neutral, 29% Democrat, 26% Republican, and 5% Unknown, whereas in 2020 the tweets were 28% Neutral, 36% Democrat, 34% Republican, 2% Unknown. Some demographic groups were much more polarized in 2020 than they were in 2016, which led to an overestimation of their support for a certain candidate in 2020 because the respective political bias correction weight was too high.

In general, we did not use follower-followee relationships because we expected that this analysis would take a prohibitively long amount of time for data collection. If we did use those relationships, we could have geolocated users who did not have geolocated tweets turned on or have their location specified on their Twitter profile. Moreover, we could have developed a method to query for all users geolocated to be in the United States and attempted to determine their political alignment using their follower-followee relationships, even if they didn't tweet about politics. Again, this would have taken a prohibitively long amount of time, but it is an idea worth exploring in the future.

Also, our sentiment analysis model does not build on the data from the 2020 election domain. This is because we don't have human-labeled tweets about the 2020 election since it is expensive to have annotators to label the data. We chose to use general sentiment analysis instead to maintain a baseline performance. If we had labeled data, we definitely could have trained a model with higher accuracy.

Another limitation of our study has to do with the accuracy of the labeling of users. We are unsure of the accuracy of assigning gender and race/ethnicity labels to users. Since we do not have any “true” labels for users, it is impossible for us to ascertain how accurate the gender and race designations for users were.

Lastly, there are many other steps we would have liked to implement with more time. For future work in this topic, consider the following recommendations:

- Use the follower-followee relationships of users to geolocate them instead of only using Twitter users with geolocated tweets turned on or location specified in their profile.
- Profile the political alignment of all users geolocated to be in the United States, and profiling them using their follower-followee relationships, instead of only using Twitter users that tweeted about politics.
- Use the follower-followee relationships of users to profile the “neutral” or “unknown” political alignment of users more accurately
- Use Twitter to estimate voter turnout rather than relying on an estimate provided by political scientists and surveyors.
- Use a more sophisticated method such as propensity weighting to compare exit polls and Twitter data to estimate the voter share of different demographic groups.
- Use a more comprehensive Twitter queries using online trend analysis with a topic model using Latent Dirichlet Allocation to find a comprehensive list of search terms.
- Use the Twitter API to get county-specific results by inputting geo-polygons for each county, which would enable us to get results for income levels, education levels, and neighborhood types such as suburban, urban, and rural designations.
- Profile users into different age groups.
- Filter out non-potential voters, bots, and trolls.
- Train a more accurate sentiment analysis model using human-labeled tweets pertaining to

US politics.

- Use the communication network to cluster more users. The base users are the ones who have a high political preference score from sentiment analysis.
- Take into account linguistic differences among groups and locations.
- Implement corrections to account for the differences in how different demographic groups in certain states tweeted differently about political candidates.
- Use time-series analysis to account for the difference in a US state's demographics and Twitter demographics.
- Use different corrections for incumbent candidates and more polarizing "tweetable" candidates like Trump.

References

1. Kahan DM. Ideology, motivated reasoning, and cognitive reflection: An experimental study. Judgment and Decision Making. 2013;8:407–424.

2. Bovet, Alexandre & Makse, Hernán. (2019). Influence of fake news in Twitter during the 2016 US presidential election. Nature Communications. 10. 10.1038/s41467-018-07761-2.

3. Shao, C., Ciampaglia, G., Varol, O., Flammini, A., & Menczer, F. (2017). The spread of misinformation by social bots.

4. C. A. Davis, O. Varol, E. Ferrara, A. Flammini, and F. Menczer. Botornot: A system to evaluate social bots. In Proceedings of the 25th International Conference Companion on World Wide Web, pages 273–274, 2016.

5. A. Malhotra, L. Totti, W. Meira Jr., P. Kumaraguru and V. Almeida, "Studying User Footprints in Different Online Social Networks," 2012 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Istanbul, 2012, pp. 1065-1070, doi: 10.1109/ASONAM.2012.184.

6. Mislove, Alan, et al. "Understanding the demographics of twitter users." Icwsm 11.5th (2011): 25.

7. Li, Linna, Michael F. Goodchild, and Bo Xu. "Spatial, temporal, and socioeconomic patterns in the use of Twitter and Flickr." Cartography and geographic information science 40.2 (2013): 61-77.

8. Rout, Dominic, et al. "Where's@ wally? a classification approach to geolocating users based on their social ties." *Proceedings of the 24th ACM Conference on Hypertext and Social Media*. 2013.

9. Backstrom, Lars, Eric Sun, and Cameron Marlow. "Find me if you can: improving

geographical prediction with social and spatial proximity." *Proceedings of the 19th international conference on World wide web*. 2010.

10. Java, Akshay, et al. "Why we twitter: understanding microblogging usage and communities." *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*. 2007.

11. Sloan, Luke. "Who Tweets in the United Kingdom? Profiling the Twitter Population Using the British Social Attitudes Survey 2015." *Social Media + Society*, Jan. 2017, doi:10.1177/2056305117698981.

12. Alowibdi, Jalal S., Ugo A. Buy, and Philip Yu. "Language independent gender classification on Twitter." *Proceedings of the 2013 IEEE/ACM international conference on advances in social networks analysis and mining*. 2013.

13. Sakaki, Shigeyuki, et al. "Twitter user gender inference using combined analysis of text and image processing." *Proceedings of the Third Workshop on Vision and Language*. 2014.

14. Mellon, Jonathan, and Christopher Prosser. "Twitter and Facebook Are Not Representative of the General Population: Political Attitudes and Demographics of British Social Media Users." *Research & Politics*, July 2017, doi:10.1177/2053168017720008.

15. Choy, Murphy, et al. "Us presidential election 2012 prediction using census corrected twitter model." *arXiv preprint arXiv:1211.0938* (2012).

16. Gayo-Avello, Daniel. "I Wanted to Predict Elections with Twitter and all I got was this Lousy Paper"--A Balanced Survey on Election Prediction using Twitter Data." *arXiv preprint arXiv:1204.6441* (2012).

17. Beauchamp, Nicholas. "Predicting and interpolating state‐level polls using Twitter textual data." American Journal of Political Science 61.2 (2017): 490-503.

18. Mellon, Jonathan, and Christopher Prosser. “Twitter and Facebook Are Not Representative of the General Population: Political Attitudes and Demographics of British Social Media Users.” *Research & Politics*, July 2017, doi:10.1177/2053168017720008.

19. Adamic, Lada A., and Natalie Glance. "The political blogosphere and the 2004 US election: divided they blog." Proceedings of the 3rd international workshop on Link discovery. 2005.

20. Rigdon, Steven E., et al. "A Bayesian prediction model for the US presidential election." American Politics Research 37.4 (2009): 700-724.

21. Anstead, Nick, and Ben O'Loughlin. "Social media analysis and public opinion: The 2010 UK general election." Journal of Computer-Mediated Communication 20.2 (2015): 204-220.

22. Shi, L.. “Predicting US Primary Elections with Twitter.” (2012).

23. Ceron, Andrea, Luigi Curini, and Stefano M. Iacus. "Using sentiment analysis to monitor electoral campaigns: Method matters—evidence from the United States and Italy." Social Science Computer Review 33.1 (2015): 3-20.

24. A multifaceted approach to social multimedia-based prediction of elections Q You, L Cao, Y Cong, X Zhang, J Luo IEEE Transactions on Multimedia 17 (12).

25. Tunggawan, Elvyna, and Yustinus Eko Soelistio. "And the winner is…: Bayesian Twitter-based prediction on 2016 US presidential election." 2016 International Conference on Computer, Control, Informatics and its Applications (IC3INA). IEEE, 2016.

26. Wicaksono, Andy Januar. "A proposed method for predicting US presidential election by analyzing sentiment in social media." *2016 2nd international conference on science in information technology (ICSITech)*. IEEE, 2016.

27. Kassraie, Parnian, Alireza Modirshanechi, and Hamid K. Aghajan. "Election Vote Share Prediction using a Sentiment-based Fusion of Twitter Data with Google Trends and Online

Polls." *DATA*. 2017.

28. Grover, Purva, et al. "Polarization and acculturation in the 2016 US presidential election: Can Twitter analytics predict changes in voting preferences?." Technological Forecasting and Social Change (2018).

29. Ahammed, Md Sohel, Md Nahid Newaz, and Arunavo Dey. "Analyzing Political Opinions and Prediction of Voting Patterns in the US Election with Data Mining Approaches." Global Journal of Computer Science and Technology (2019).

30. Sanders, Eric, and Antal van den Bosch. "Optimising Twitter-based Political Election Prediction with Relevance andSentiment Filters." Proceedings of The 12th Language Resources and Evaluation Conference. 2020.

31. Shirodkar, Madhura et al. "Election result prediction using Sentiment Analysis." (2020).

32. Mohsen Sayyadiharikandeh, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. 2020. Detection of Novel Social Bots by Ensembles of Specialized Classifiers.

33. Miranda Filho, Renato, Jussara M. Almeida, and Gisele L. Pappa. "Twitter population sample bias and its impact on predictive outcomes: a case study on elections." *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, 2015.

34. Sentiment Analysis of Twitter Data: A Survey of Techniques. *International Journal of Computer Applications* 139(11):5-15, April 2016

35. Twitter Population Sample Bias and its impact on predictive outcomes: a case study on elections

36. Using sentiment analysis to define twitter political users' classes and their homophily during the 2016 American presidential election

37. Ms. Farha Nausheen et al., “Sentiment Analysis to Predict Election Results Using Python”, Proceedings of the Second International Conference on Inventive Systems and Control (ICISC 2018).

38. Ms. Farha Nausheen et al., “Sentiment Analysis to Predict Election Results Using Python”, Proceedings of the Second International Conference on Inventive Systems and Control (ICISC 2018)

39. A multifaceted approach to social multimedia-based prediction of elections Q You, L Cao, Y Cong, X Zhang, J Luo IEEE Transactions on Multimedia 17 (12)

40. Vepsäläinen, Tapio, Hongxiu Li, and Reima Suomi. "Facebook likes and public opinion: Predicting the 2015 Finnish parliamentary elections." *Government Information Quarterly* 34.3 (2017): 524-532.

41. Gayo-Avello, Daniel. "A meta-analysis of state-of-the-art electoral prediction from Twitter data." *Social Science Computer Review* 31.6 (2013): 649-679

42. Beauchamp, Nicholas. "Predicting and interpolating state‐level polls using Twitter textual data." *American Journal of Political Science* 61.2 (2017): 490-503.

43. Ginsberg, Jeremy, et al. "Detecting influenza epidemics using search engine query data." *Nature* 457.7232 (2009): 1012-1014.

44. Lau, Jey Han, Nigel Collier, and Timothy Baldwin. "On-line trend analysis with topic models:# twitter trends detection topic model online." *Proceedings of COLING 2012*. 2012.

45. Tumasjan, Andranik, et al. "Predicting elections with twitter: What 140 characters reveal about political sentiment." *Fourth international AAAI conference on weblogs and social media*. 2010.

46. Jürgens, Pascal, Andreas Jungherr, and Harald Schoen. "Small worlds with a difference: New

gatekeepers and the filtering of political information on Twitter." *Proceedings of the 3rd international web science conference*. 2011.

47. Gayo Avello, Daniel, Panagiotis T. Metaxas, and Eni Mustafaraj. "Limits of electoral predictions using twitter." *Proceedings of the fifth international AAAI conference on weblogs and social media*. Association for the Advancement of Artificial Intelligence, 2011.

48. Skoric, Marko, et al. "Tweets and votes: A study of the 2011 singapore general election." *2012 45th hawaii international conference on system sciences*. IEEE, 2012.

49. Sang, Erik Tjong Kim, and Johan Bos. "Predicting the 2011 dutch senate election results with twitter." *Proceedings of the workshop on semantic analysis in social media*. 2012.

50 O'Connor, Brendan, et al. "From tweets to polls: Linking text sentiment to public opinion time series." *Tepper School of Business* (2010): 559.

51. Ceron, Andrea, Luigi Curini, and Stefano M. Iacus. "Using sentiment analysis to monitor electoral campaigns: Method matters—evidence from the United States and Italy." Social Science Computer Review 33.1 (2015): 3-20.

52. Livne, Avishay, et al. "The party is over here: Structure and content in the 2010 election." *ICWSM* 11.2011 (2011): 17-21.

53. Metaxas, Panagiotis T., Eni Mustafaraj, and Dani Gayo-Avello. "How (not) to predict elections." *2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*. IEEE, 2011.

54. Gayo Avello, Daniel, Panagiotis T. Metaxas, and Eni Mustafaraj. "Limits of electoral predictions using twitter." *Proceedings of the fifth international AAAI conference on weblogs and social media*. Association for the Advancement of Artificial Intelligence, 2011.

55. Mellon, Jonathan, and Christopher Prosser. "Twitter and Facebook are not representative of

the general population: Political attitudes and demographics of British social media users." *Research & Politics* 4.3 (2017): 2053168017720008.

56. Huang, Yuan, et al. "Understanding US regional linguistic variation with Twitter data analysis." *Computers, environment and urban systems* 59 (2016): 244-255.

57. Conover, Michael D., et al. "Predicting the political alignment of twitter users." *2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing*. IEEE, 2011.

58. Wang, Hao, et al. "A system for real-time twitter sentiment analysis of 2012 us presidential election cycle." *Proceedings of the ACL 2012 system demonstrations*. 2012.

59. Yang, Kai-Cheng, Pik-Mai Hui, and Filippo Menczer. "How Twitter Data Sampling Biases US Voter Behavior Characterizations." *arXiv preprint arXiv:2006.01447* (2020).

60. Bethlehem, Jelke. "Selection bias in web surveys." *International Statistical Review* 78.2 (2010): 161-188.