Election Discussion on Twitter: Topics, Sentiment, and Predictions

Xinran Wang
Columbia University
xw2418@columbia.edu

Yijia Zhou
Columbia University
yz2859@columbia.edu

Abstract

This paper examines a sample of daily Twitter discussion on U.S. president election 2016. It includes an exploration of Twitter trends and patterns across states, political topics, candidates and parties. We also performed a text analysis on the sample collection of tweets to determine to what extent they are similar to each other. Last, we attempted to make a prediction model, which uses tweet number, discussion pattern, and tweeting time frame to predict primary results. It is a preliminary analysis of tweets that aims to improve understanding of political discussion and potential correlation with real-life results, and both magnitude and attitude of tweets are found to be related to the outcome. Our result demonstrates the practicality of using social media to understand public opinion and trend on election polling.

Keywords

Twitter; Election 2016; U.S. primaries

Introduction

The representation of 2016 U.S election on Twitter exhibits popular political and social trend in the perspective of network media. As the widely used social platform, Twitter provides abundant information to detect the trend in elections and sheds some light to the insight of the correlations. We start with visualization to process the massive Twitter data and obtain an overall pattern of the descriptive statistics. Using cross filter, the daily frequency, most discussed topics, share of each candidate and geographic information are displayed. To explore the relationships between factors of interest more precisely, we use text analysis and make predictions on the primary results. Text analysis captures the tweets involving each candidate, examines the topics and analyzes the sentiments. Though sarcasm, humor and other subtle use of English vocabulary are hard to detect, the result demonstrates the variations in topic distribution and frequent terms when Twitter user talk about major candidates. Predictions are based on the volumes, retweet frequencies, and sentiment categorization. Direction of influence of independent variables differs by each candidate. The time correlation in the tweets also allows us to make appropriate forecast with time trend.

Data Collection and Cleaning

For the purpose of this study, tweets were collected daily from March 15 to April 15, 2016, using keywords of major presidential election candidates' names. Used keywords included: "Hillary Clinton", "Bernie Sanders", "Jeb Bush", "Marco Rubio","Ted Cruz", "Donald Trump", "John Kasich", "Ben Carson", "Chris Christie", "Carly Fiorina", "Hillaryclinton", "Jebbush", "Berniesanders", "Marcorubio", "Chrischristie", "Bencarson", "Johnkasich", "Donaldtrump", "Tedcruz", "CarlyFiorina","#feelthebern","Clinton", "@hillaryclinton", "@berniesanders","@sensanders", "@martinomalley","@realdonaldtrump","@tedcruz","@johnkasich","@marcorubio","@chrischristie","@realbencarson",and "@jebbush".

A sample of tweets was generated so that it contained 5,000 tweets per day. While we acknowledge the fact that 5,000 tweets per day was a relatively small number as compared to the vast amount of tweets created each day, we are limited by technical difficulties that constrained our ability to process larger files. Then, since this paper aims at detecting major patterns of political discussion on Twitter, we subsetted the data to include only tweets that concerned the four leading candidates, two from each party: Hillary Clinton, Bernie Sanders, Donald Trump, and Ted Cruz. That left us with 13,4037 tweets in total to work with.

Like stated in the introduction, we are primarily concerned about these questions: what are people talking about when they talk about leading candidates of Election 2016 Twitter? When are they tweeting? And where? How are they feeling when tweeting? With that in mind, we further processed the data to distinguish tweets by candidate, by party, and also by geographic location of Twitter users. Since only a very limited number of tweets(less than 1%) contained the real-time geographic location of user, for geographic location we used the location in user profile instead. This may compromise the validity of our study at geographic distribution of tweets, but we believe it is a necessary one that provided enough data points to make any generalizable conclusion.

Text Analysis

With processed data, we started by performing a text analysis, which included topic detection by LDA, topic classification, the distribution of topics, semantic analysis, and a comparison of similarity across tweets by candidate and by party.

Topic Detection

Before actually categorizing tweets into different topics, we performed a Latent Dirichlet Allocation on available data to explore potential topics of discussion. The LDA algorithm was chosen because it, unlike humans, is unaffected by prior beliefs and bias and therefore more likely to be more objective. Also, LDA is able to capture and distinguish different uses of a certain term based on context, making it an effective method for topic detection. We set the assumed number of topics to be ten so that the number is large enough to fully capture the complexity of discussion, but not too large to be less meaningful for analysis.

As shown in Table 1 and 2, we are unsurprised to find that the majority of terms are related to election behaviors, results, and news closely related to election candidates, such as Cruz's sex scandal and Clinton's leaked emails. States holding primaries during that time frame such as Wisconsin and Utah appeared frequently in the term list as a result. For the same reason, major news outlets such as Fox News and CNN also appeared multiple times. Also, there is an abundance of supporter and/or haters' hashtags, including Feelthebern, makeamericagreatagain, and choosecruz. Notably Donald Trump also had a hater hashtag, "neverTrump", that appears in the term list with AlwaysTrump, indicating the polarized opinion about the candidate on Twitter. Apart from these obvious signs for candidate preference, there are also words like "rally", "believe" and "support" that were probably used by supporters to boost the morale before primaries. The swearing word "fuck" also appeared in the list of terms, which made us curious about the use of swearing words in Twitter discussion and we would explore that further in the next section.

The distinction between automatically detected topics, however, is not that clear. Many terms appeared in multiple lists, if not all of them, and there is an observable overlap among different topic categories. First, all four candidates' names ranked quite high in the list, and the majority of terms that repeatedly appeared belongs to the vocabulary of politics, such as "gop", "vote" and "delegate". This high level of overlap is surprising but understandable, and it can be explained by the vast amount of retweeting messages of news reports on Twitter. While these kind of tweets do not contain much actual content, they tend to be retweeted most often, which affected the classification of LDA. Therefore, a manual selection of keywords for topic classification would be used to control the effect of Twitter nature.

Still, there are some noticeable difference between topics. For instance, Topic 6 is likely to be about Bernie Sanders. More accurately, it is mostly generated by Sanders' supporters, as shown by the appearance of terms like "demdebate", "feelthebern", "stillsanders" and "birdiesanders". Topic 1, on the other hand, probably contains the discussion about

Wisconsin primary results, as illustrated by the highly ranked “wisconsin”, “foxnews”, and “trumptrain”.

Overall, while the Latent Dirichlet Allocation did not directly generate classified topics for us, it worked as an efficient way to explore the vast collection of texts on Twitter, and helped us to determine which topic categories we should start from. More specifically, the result led us to believe it is necessary to create a topic category that is exclusively about election news and results, so that we could observe actual tweet discussion about politics without being masked by the overwhelming proportion of news retweets.

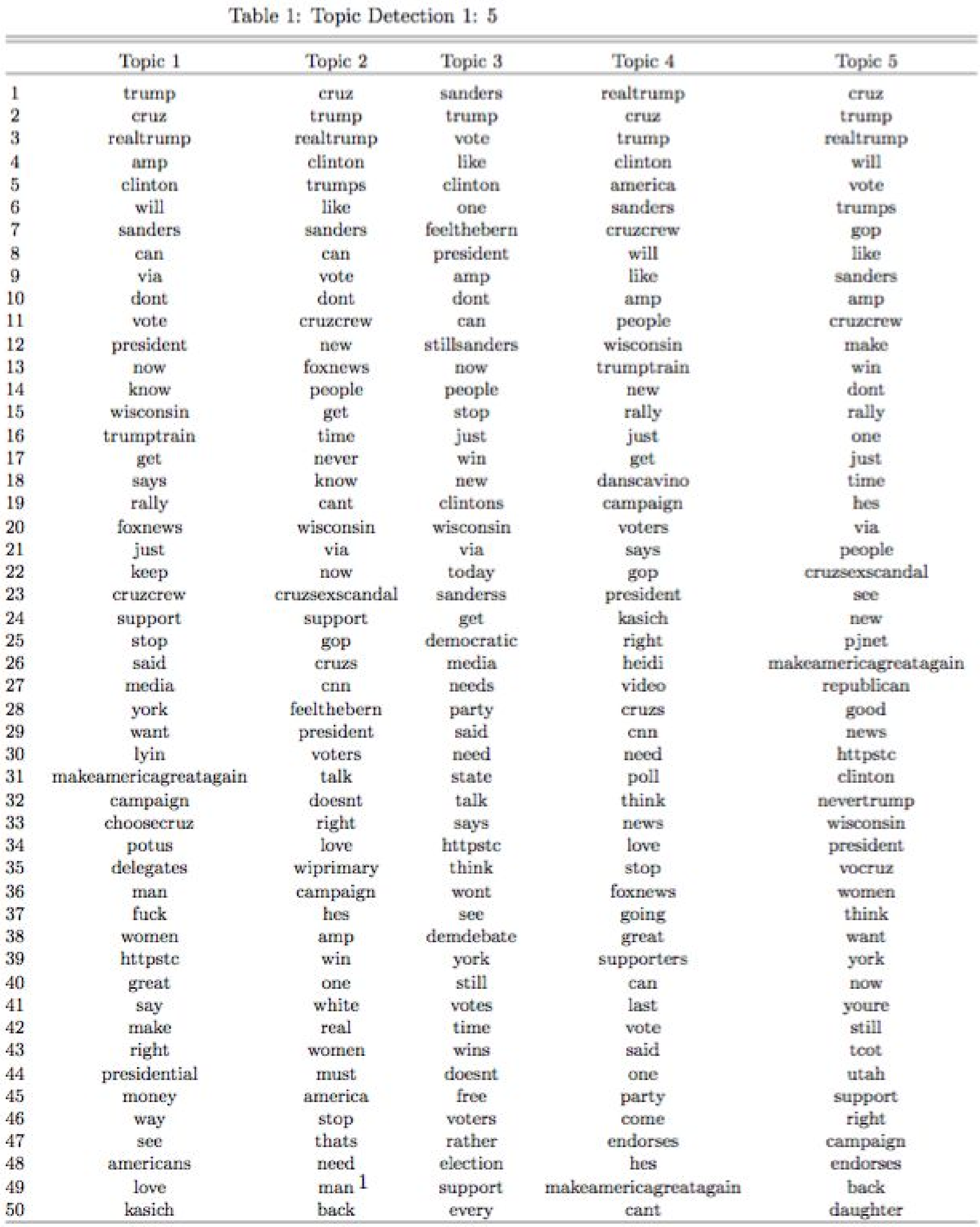


Table 1: Topic Detection 1: 5

	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5
1	trump	cruz	sanders	realtrump	cruz
2	cruz	trump	trump	cruz	trump
3	realtrump	realtrump	vote	trump	realtrump
4	amp	clinton	like	clinton	will
5	clinton	trumps	clinton	america	vote
6	will	like	one	sanders	trumps
7	sanders	sanders	feelthebern	cruzcrew	gop
8	can	can	president	will	like
9	via	vote	amp	like	sanders
10	dont	dont	dont	amp	amp
11	vote	cruzcrew	can	people	cruzcrew
12	president	new	stillsanders	wisconsin	make
13	now	foxnews	now	trumptrain	win
14	know	people	people	new	dont
15	wisconsin	get	stop	rally	rally
16	trumptrain	time	just	just	one
17	get	never	win	get	just
18	says	know	new	danscavino	time
19	rally	cant	clintons	campaign	hes
20	foxnews	wisconsin	wisconsin	voters	via
21	just	via	via	says	people
22	keep	now	today	gop	cruzsexscandal
23	cruzcrew	cruzsexscandal	sanderss	president	see
24	support	support	get	kasich	new
25	stop	gop	democratic	right	pjnet
26	said	cruzs	media	heidi	makeamericagreatagain
27	media	cnn	needs	video	republican
28	york	feelthebern	party	cruzs	good
29	want	president	said	cnn	news
30	lyin	voters	need	need	httpstc
31	makeamericagreatagain	talk	state	poll	clinton
32	campaign	doesnt	talk	think	nevertrump
33	choosecruz	right	says	news	wisconsin
34	potus	love	httpstc	love	president
35	delegates	wiprimary	think	stop	vocruz
36	man	campaign	wont	foxnews	women
37	fuck	hes	see	going	think
38	women	amp	demdebate	great	want
39	httpstc	win	york	supporters	york
40	great	one	still	can	now
41	say	white	votes	last	youre
42	make	real	time	vote	still
43	right	women	wins	said	tcot
44	presidential	must	doesnt	one	utah
45	money	america	free	party	support
46	way	stop	voters	come	right
47	see	thats	rather	endorses	campaign
48	americans	need	election	hes	endorses
49	love	man[1]	support	makeamericagreatagain	back
50	kasich	back	every	cant	daughter

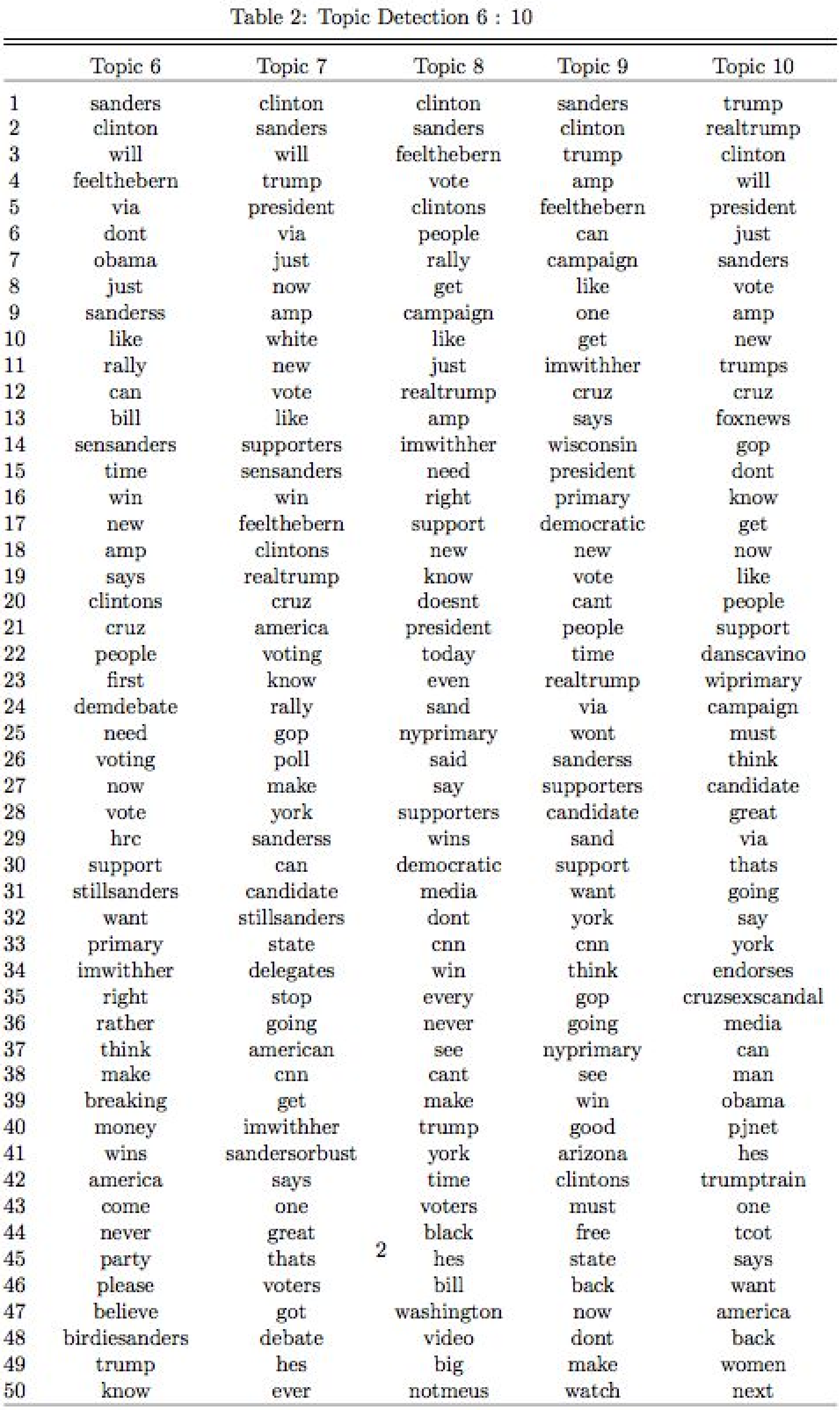


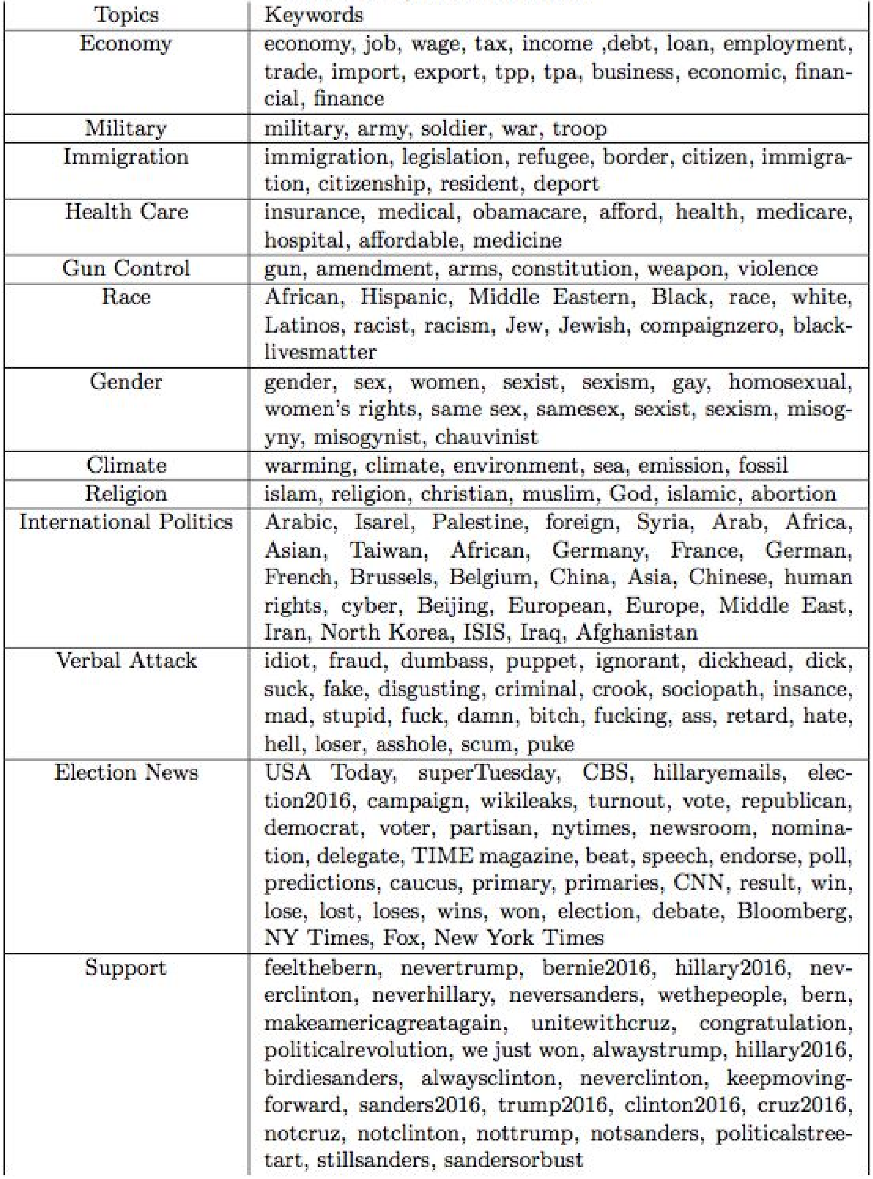
Table 2: Topic Detection 6 : 10

	Topic 6	Topic 7	Topic 8	Topic 9	Topic 10
1	sanders	clinton	clinton	sanders	trump
2	clinton	sanders	sanders	clinton	realtrump
3	will	will	feelthebern	trump	clinton
4	feelthebern	trump	vote	amp	will
5	via	president	clintons	feelthebern	president
6	dont	via	people	can	just
7	obama	just	rally	campaign	sanders
8	just	now	get	like	vote
9	sanderss	amp	campaign	one	amp
10	like	white	like	get	new
11	rally	new	just	imwithher	trumps
12	can	vote	realtrump	cruz	cruz
13	bill	like	amp	says	foxnews
14	sensanders	supporters	imwithher	wisconsin	gop
15	time	sensanders	need	president	dont
16	win	win	right	primary	know
17	new	feelthebern	support	democratic	get
18	amp	clintons	new	new	now
19	says	realtrump	know	vote	like
20	clintons	cruz	doesnt	cant	people
21	cruz	america	president	people	support
22	people	voting	today	time	danscavino
23	first	know	even	realtrump	wiprimary
24	demdebate	rally	sand	via	campaign
25	need	gop	nyprimary	wont	must
26	voting	poll	said	sanderss	think
27	now	make	say	supporters	candidate
28	vote	york	supporters	candidate	great
29	hrc	sanderss	wins	sand	via
30	support	can	democratic	support	thats
31	stillsanders	candidate	media	want	going
32	want	stillsanders	dont	york	say
33	primary	state	cnn	cnn	york
34	imwithher	delegates	win	think	endorses
35	right	stop	every	gop	cruzsexscandal
36	rather	going	never	going	media
37	think	american	see	nyprimary	can
38	make	cnn	cant	see	man
39	breaking	get	make	win	obama
40	money	imwithher	trump	good	pjnet
41	wins	sandersorbust	york	arizona	hes
42	america	says	time	clintons	trumptrain
43	come	one	voters	must	one
44	never	great	black	free	tcot
45	party	thats	hes	state	says
46	please	voters	bill	back	want
47	believe	got	washington	now	america
48	birdiesanders	debate	video	dont	back
49	trump	hes	big	make	women
50	know	ever	notmeus	watch	next

Topic Classification

When it comes to create topic categories, we combined results of the LDA algorithm with human experience. As shown in Table 3, we sorted the tweets into 13 different topic categories. Topics and keywords are selected based on frequency, and with each step of topic categorization, we also examined texts of unclassified tweets to improve our keywords selection. While other topic categories are relatively straightforward, election, support, and verbal attacks are three categories worth explaining.

Table 3: Keywords of Topics



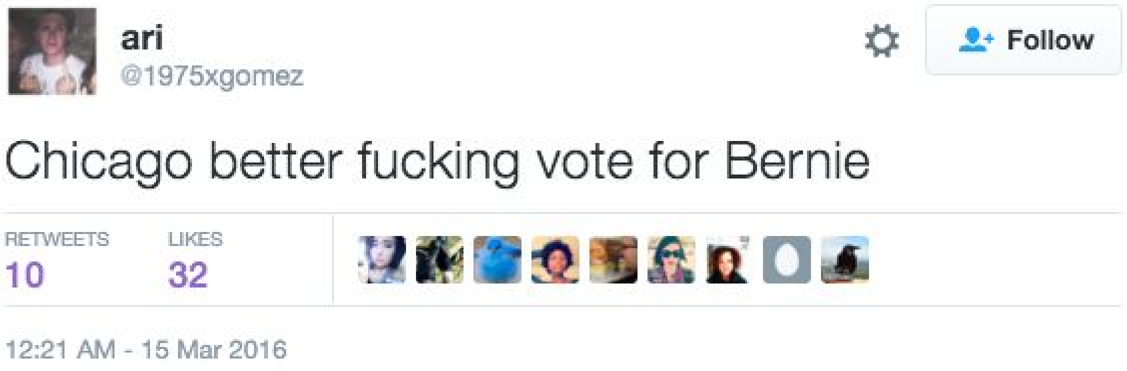
Topics	Keywords
Economy	economy, job, wage, tax, income ,debt, loan, employment, trade, import, export, tpp, tpa, business, economic, financial, finance
Military	military, army, soldier, war, troop
Immigration	immigration, legislation, refugee, border, citizen, immigration, citizenship, resident, deport
Health Care	insurance, medical, obamacare, afford, health, medicare, hospital, affordable, medicine
Gun Control	gun, amendment, arms, constitution, weapon, violence
Race	African, Hispanic, Middle Eastern, Black, race, white, Latinos, racist, racism, Jew, Jewish, compaignzero, blacklivesmatter
Gender	gender, sex, women, sexist, sexism, gay, homosexual, women's rights, same sex, samesex, sexist, sexism, misogyny, misogynist, chauvinist
Climate	warming, climate, environment, sea, emission, fossil
Religion	islam, religion, christian, muslim, God, islamic, abortion
International Politics	Arabic, Isarel, Palestine, foreign, Syria, Arab, Africa, Asian, Taiwan, African, Germany, France, German, French, Brussels, Belgium, China, Asia, Chinese, human rights, cyber, Beijing, European, Europe, Middle East, Iran, North Korea, ISIS, Iraq, Afghanistan
Verbal Attack	idiot, fraud, dumbass, puppet, ignorant, dickhead, dick, suck, fake, disgusting, criminal, crook, sociopath, insance, mad, stupid, fuck, damn, bitch, fucking, ass, retard, hate, hell, loser, asshole, scum, puke
Election News	USA Today, superTuesday, CBS, hillaryemails, election2016, campaign, wikileaks, turnout, vote, republican, democrat, voter, partisan, nytimes, newsroom, nomination, delegate, TIME magazine, beat, speech, endorse, poll, predictions, caucus, primary, primaries, CNN, result, win, lose, lost, loses, wins, won, election, debate, Bloomberg, NY Times, Fox, New York Times
Support	feelthebern, nevertrump, bernie2016, hillary2016, neverclinton, neverhillary, neversanders, wethepeople, bern, makeamericagreatagain, unitewithcruz, congratulation, politicalrevolution, we just won, alwaystrump, hillary2016, birdiesanders, alwaysclinton, neverclinton, keepmovingforward, sanders2016, trump2016, clinton2016, cruz2016, notcruz, notclinton, nottrump, notsanders, politicalstreetart, stillsanders, sandersorbust

Twitter is a platform for discussion as well as news dissemination, and in the case of election discussion, it is only natural that we found a substantial amount of tweets to be retweets of election-related news: primary results, debates, and also highly related news such as the leak of Hillary Clinton's emails. Many of them contained very few actual content, but were merely a retweet of major news sites. Therefore, we concluded that it would be helpful to contain tweets that were more about spreading the news than discussing the results in one single category. And for those tweets which actually discussed military, gender, international politics etc when retweeting a news site, they have an equal possibility to be categorized into other topic categories.

Support(orNot) is another category of tweets that appeared frequently but is also likely to contain few actual discussion. These tweets are mainly support messages from vocal supporters of candidates, who are inclined to tweet with signature hashtags such as #feelthebern, #AlwaysTrump, and #Hillary2016. We also included anti-candidate hashtags such as #NeverTrump, because they are also a declaration of attitude towards candidates, and also likely to contain minimum content. Support(orNot) tweets tended to appear most frequently around primaries for morale support.

Verbal attack obviously contained a list of swearing words that are less suitable in presentation, but ultimately essential if we want a whole picture of Twitter sphere. These are tweets that . It is worth noting that having more verbal attack tweets does not necessarily mean that the candidate is attacked most often. It simply means that people are more likely to swear when talking about him/her. For example, this tweet below is obviously a support message for Bernie Sanders, not an attack on him.

Sample Tweet 1



ari
@1975xgomez

Chicago better fucking vote for Bernie

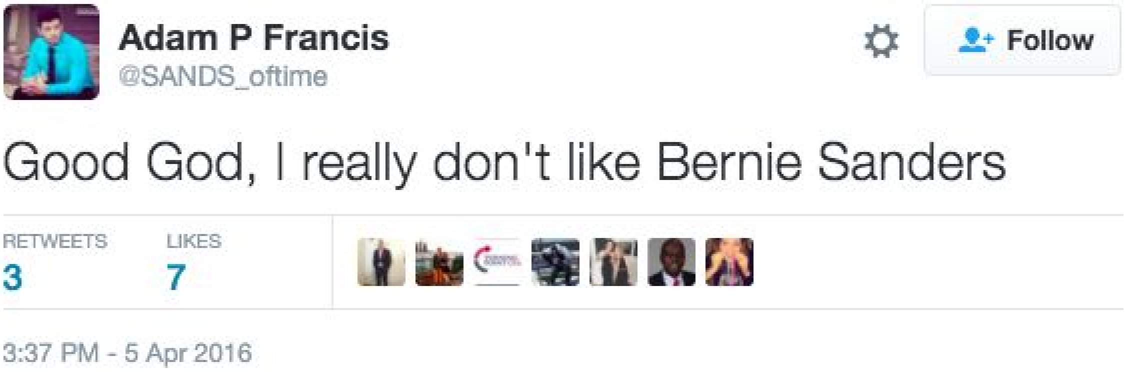
RETWEETS 10
LIKES 32

12:21 AM - 15 Mar 2016

As discussed above, we are aware that some tweets may be classified into more than one category, but we trust the algorithm to randomly assign a tweet into one category when several possible categories existed.

Unfortunately, it is hard for the program to detect sarcasm, humor, and other subtle use of English vocabulary, which may also lead to misclassification. For instance, when people tweeted "God", they could be exclaiming, like shown in the tweet below, instead of talking about religion. Similarly, when people mentioned "war" on Twitter, it's also possible that they were discussing "the presidential war" instead of an actual war. But these are all risks that we had to take as researchers, and overall the classification method had a good performance.

Sample Tweet 2



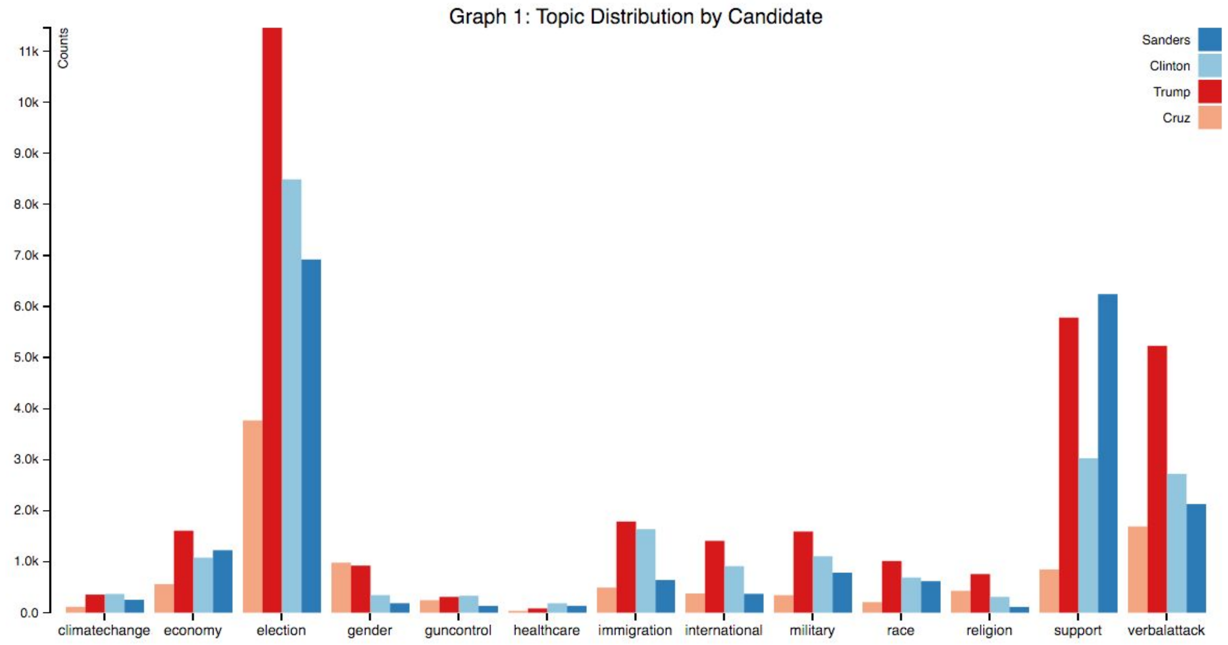
Adam P Francis
@SANDS_oftime

Good God, I really don't like Bernie Sanders

RETWEETS 3 LIKES 7

3:37 PM - 5 Apr 2016

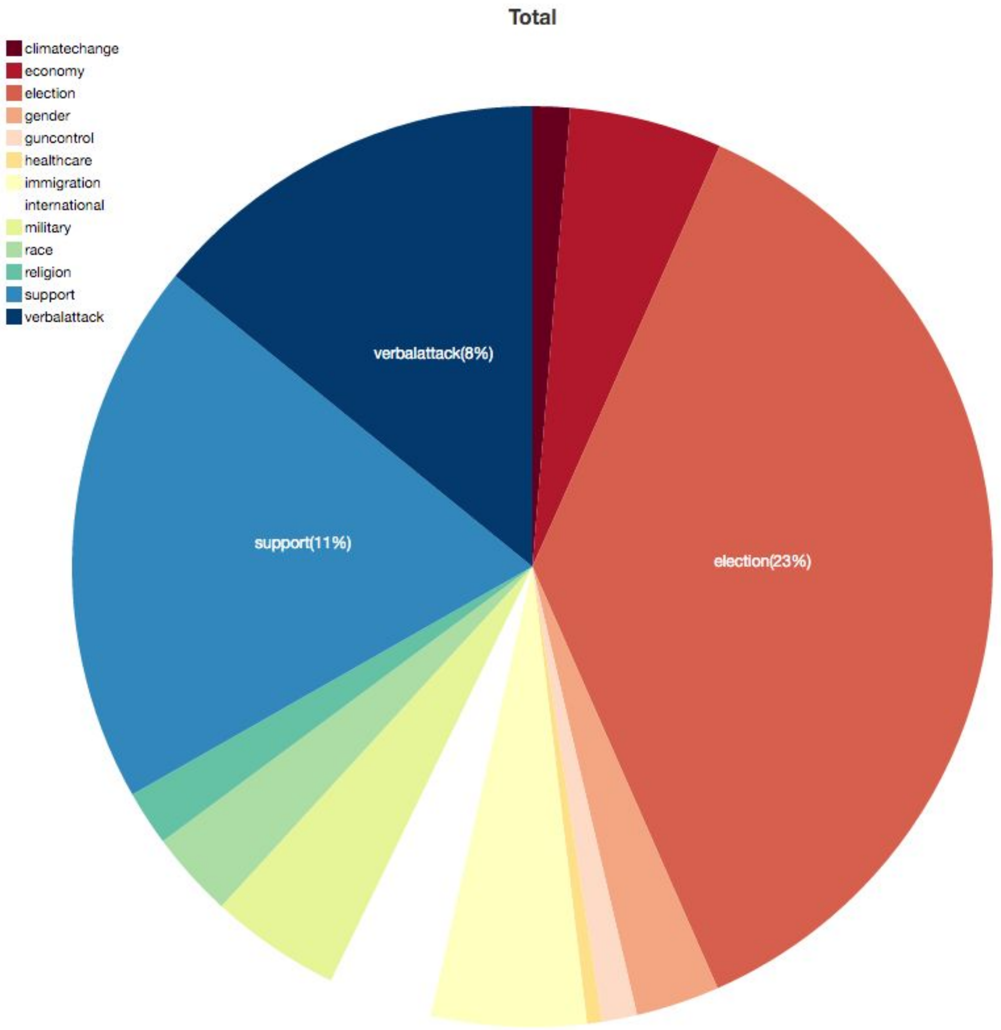
Topic Distribution



Graph 1 is an illustration of overall distribution of topics by candidate from March 15 to April 15, 2016. Ranked from high to low, the most frequently mentioned topics across all candidates are: election, support, verbal attack, immigration, economy, military, international politics, race, gender and sexuality, religion, climate change, gun control and health care.

As illustrated by Graph 2, election news made up almost ¼ of all tweets, support(orNot) tweets account for over 10% of total tweets, and for every 100 tweets, there are 8 tweets that include swearing words. The rest of topic categories ranked from high to low by proportion as such: immigration(5.47%), economy(5.36%), military(4.59%), international politics(3.68%), race(3.03%), gender & sexuality(2.92%), religion(1.93%), climate change(1.31%), gun control(1.22%), and health care(0.53%). This overall distribution of topics confirmed our assumption after performing the LDA analysis that election news and support/hate messages constituted the bulk of election discussion on Twitter. The distribution of remaining topics is to some extent unexpected: while immigration and economy are two topics that attracted many attention and understandably ranked high in frequency, military ranked higher than we expected, and discussion about health care surprisingly only made up 0.53% of the entire discussion. It is possible that our choice of keywords for military affected the result: “war” was used as one of the keywords for military topic, but Twitter users could be talking metaphorically about the election war, just as tweets which contained “race” could be about “the presidential race”, not about race as in race and ethnicity. However, after a closer examination of tweets classified as military, we are able to conclude that while the suspected situation occurred, the majority of military tweets is still correctly classified. Namely, military is a topic brought up quite frequently by Twitter users when they are discussing election 2016 candidates. Similarly, as surprising as it is, health care is not very relevant in political discussion as illustrated by currently available data. It could be influenced by our relatively small sample size, and we would in the future perform an analysis using larger dataset with enough technical support.

Graph 2: Overall Topic Distribution



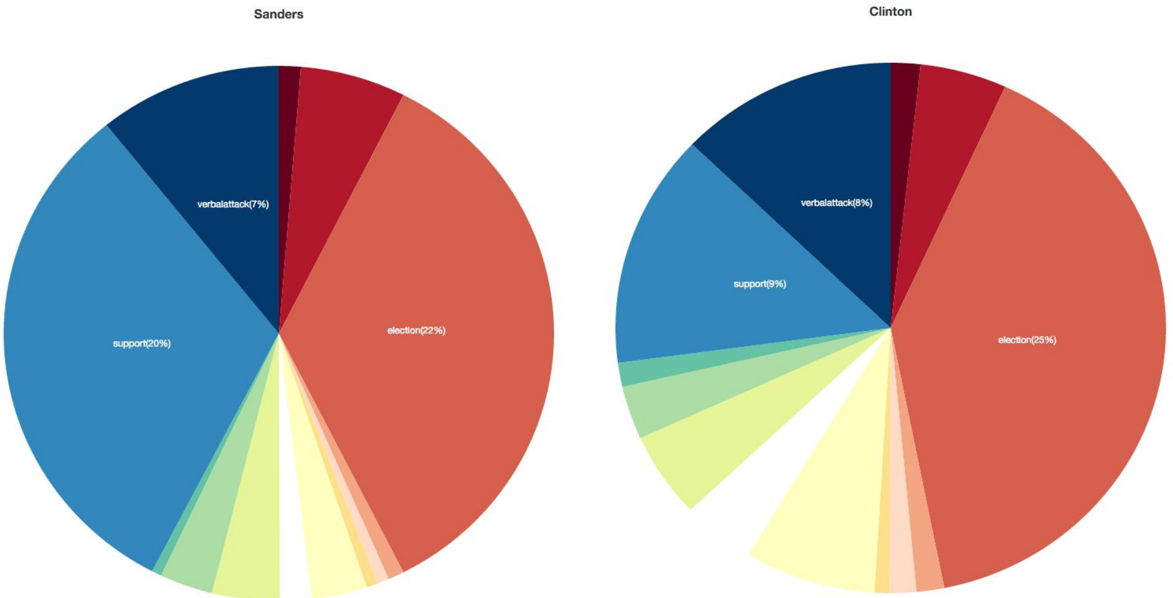
[1]

It is worth noting that the order of candidates' distribution roughly remained stable across topics: Trump is the most frequently discussed candidate for all topics except climate change, health care, gun control, and gender & sexuality, and even for these three topics, he ranked the second and had a small margin(not more than 100 tweets less than the first candidate). Cruz, on the other hand, was rarely mentioned as compared to other candidates, since he always ranked the lowest in frequency across topics except gender and sexuality, and it was by a large margin. It is quite common that

[1] Link for shinyapp: https://xinranwang.shinyapps.io/visfinal/
Github: https://xinranwang.shinyapps.io/visfinal/

Cruz only had a third of mentioning of Trump's when they are mentioned in the same topic. Sanders and Clinton, however, had a more diverse distribution of topics. For instance, Sanders exceeds Trump to be the most frequently mentioned candidate in support, but only ranked the third in election, verbal attack, and had less than half tweets as compared to Clinton in the immigration topic. Clinton, on the other hand, is frequently mentioned when it comes to election news, immigration and international politics, but does not attract much attention in economy and gender&sexuality. Nevertheless, it should be noted that we are discussing order of candidates ranked by frequency here, which reflects the crude number of tweets, not the proportion. Therefore, Trump would rank the highest in most topics because he is the most controversial candidate and thus most frequently discussed on Twitter. To obtain a more comprehensive view of each candidate's topic distribution, it is necessary for us to examine the proportional distribution as illustrated by the pie charts below.

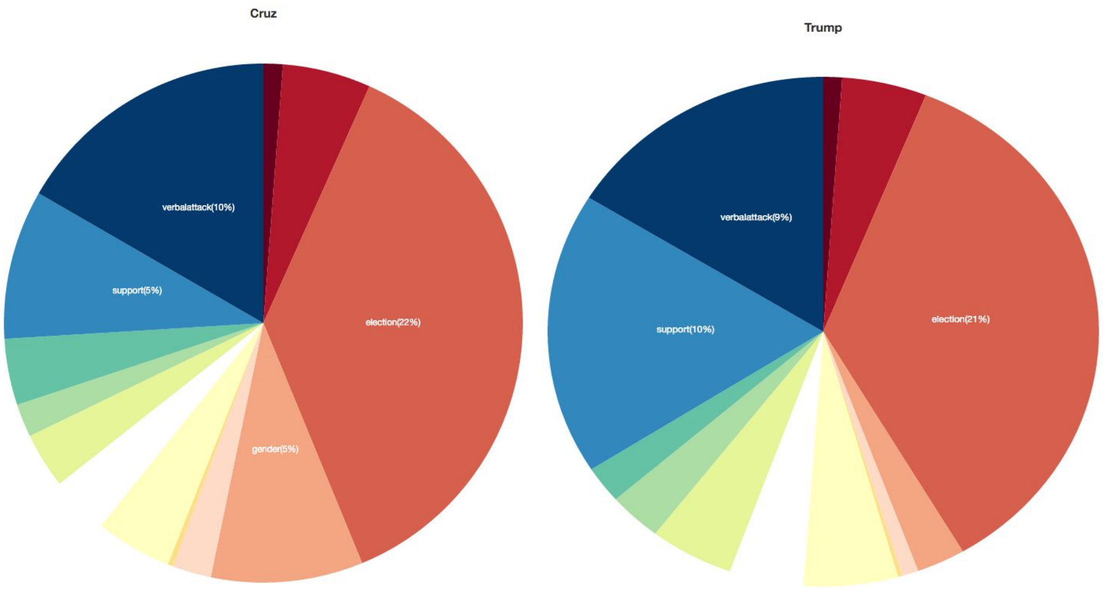
Graph 3: Topic Distribution for Sanders and Clinton



For democratic candidates, election tweets ranked the 1st in proportion for both Sanders(35.0%) and Clinton(40.1%). Sanders, however, has 31.6% support(orNot) tweets, but Clinton only has 14.3%, which reflects their difference in popularity among Twitter users. They don't differ much in verbal attack(Sanders 10.8%, Clinton 12.8%), economy(Sanders 6.20%, Clinton 5.01%), race(Sanders 3.13%, Clinton

3.25%), and health care(Sanders 0.68%, Clinton 0.87%). However, Clinton has about twice the proportion of tweets in international politics(4.30%) as compared to Sanders(1.87%), and more than twice in immigration(Sanders 3.25%, Clinton 7.72%). The high percentage of international politics for Hillary Clinton can be attributed to her past role as secretary of state, the leaked emails, and also her statement on Brussel's attack. She also has a notably higher percentage of tweets for gender & sexuality(Sanders 0.94%, Clinton 1.63%), religion(Sanders 0.57%, Clinton 1.46%), gun control(Sanders 0.68%, Clinton 1.57%), climate change(Sanders 1.28%, Clinton 1.74%) and military(Sanders 3.97%, Clinton 5.22%). In other words, Sanders has a very active group of supporters on Twitter, which constitutes 31.6% of tweets which mentioned him by name, while Clinton has less supporter messages but is mentioned more frequently with actual political discussions.

Graph 4: Topic Distribution for Cruz and Trump



For Republican candidates, it is obvious that Trump has a significantly larger volume of tweets than Cruz, often by three times and more. However, when it comes to proportional distribution, they are not that different from each other. Cruz and Trump have a similar number of percentage in election(Trump 35.5%, Cruz 37.3%), verbal attack(Trump 16.2%, Cruz 16.7%), immigration(Trump 5.53%, Cruz 4.89%), economy(Trump 4.97%, Cruz 5.54%), and climate change(Trump 1.10%, Cruz 1.14%). However, Trump has 17.9% tweets classified as support(orNot), but Cruz only has 8.40%

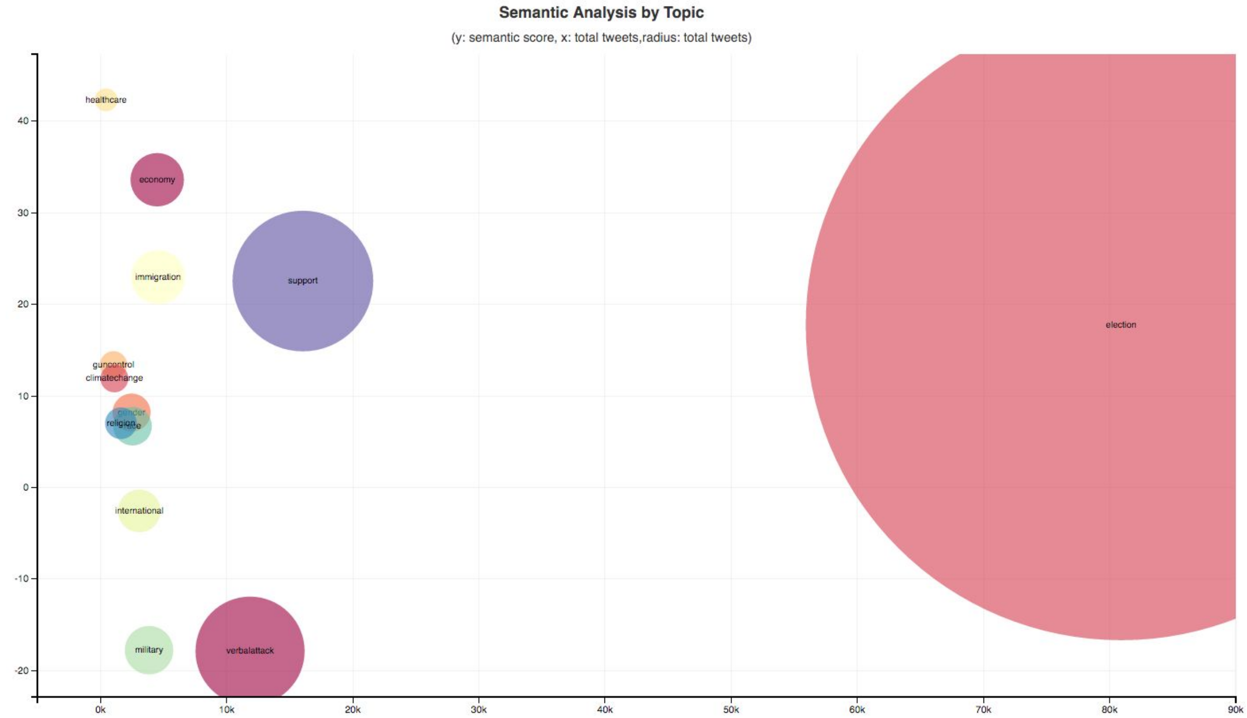
of tweets that's obviously generated by supporters or haters. This difference reflects their wide gap of popularity, or, to be more precise, controversy among the Twitter user population. Trump is also more often mentioned in military(Trump 4.92%, Cruz 3.41%), international politics(Trump 5.36%, Cruz 3.74%) and race(Trump 3.14%, Cruz 2.05%). These differences are within our expectations, because Trump tends to be very vocal about his opinion on international affairs, and is known for his controversial view on race and ethnicity. Cruz, on the other hand, has a higher percentage of tweets in gender & sexuality(Trump 2.86%, Cruz 9.72%), religion(Trump 2.34%, Cruz 4.25%) and gun control(Trump 0.96%, Cruz 1.14%). It's worth mentioning that the exceptionally high percentage of gender & sexuality tweets for Ted Cruz is a result of his sex scandal, as the wide use of hashtag "cruzsexscandal" affected the percentage. HIs high percentage in religion can be attributed to his speech on protecting the Muslim neighborhood on March 22, 2016.

Overall, the four major candidates differ greatly in volume for each topic, but not so much in the proportion distribution of topics. All four candidates has around 30% of tweets classified as election, and the support(orNot) category reflects their difference in Twitter popularity: Sanders and Trump are given most attention, followed by Clinton, but Cruz has a small number of supporters as compared to others. Democratic candidates have a smaller percentage of verbal attacks tweets in general than their Republican counterparts. For remaining topics, most of the difference among candidates' percentages can be explained by their different focus on each topics in politics.

Semantic Analysis

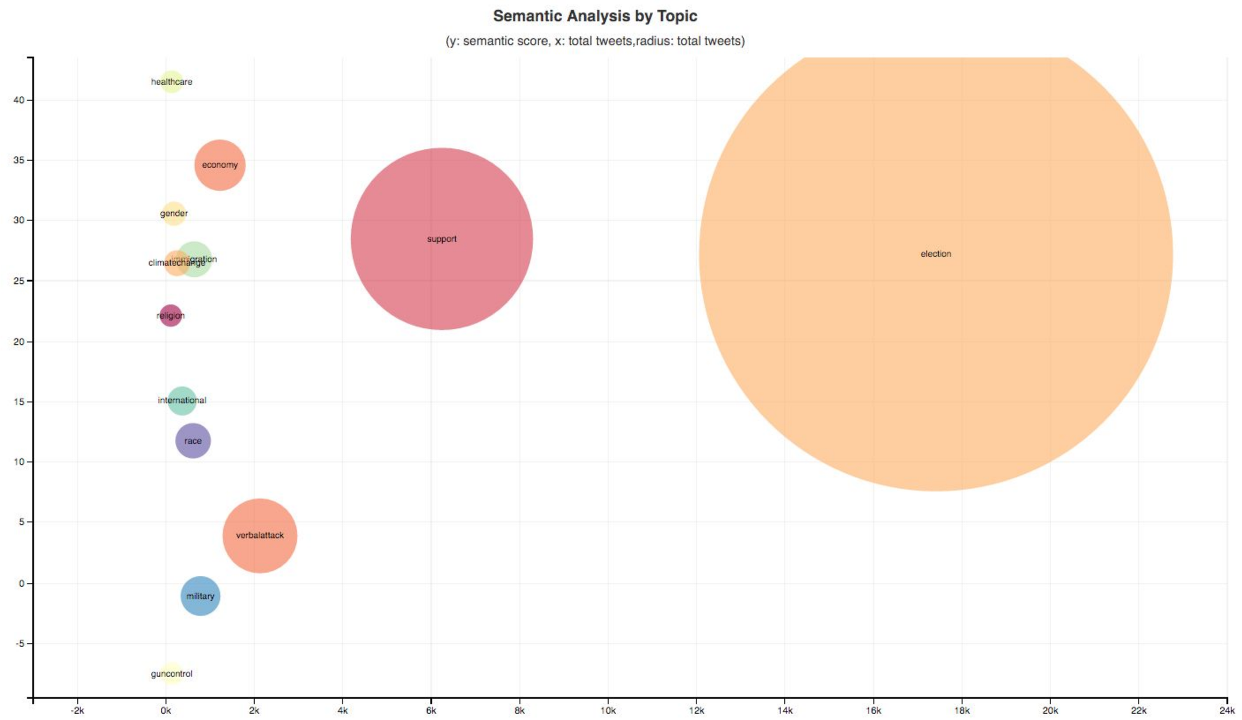
After tweets are classified by party, candidate, and topic, we are interested in the feelings expressed by tweets in different categories. To analyze that, we performed a semantic analysis using the lexicon file provided by Neal Caren and Pablo Barbera, and rescaled the semantic score by 100 times(range : [-100, 100]) for the clarity of visualization. The higher the semantic score is, the more positive the tweets are. Then, we plot the sentiment score of each topic by candidate as a bubble overlay chart.

Graph 5: Overall Semantic Score by Topic

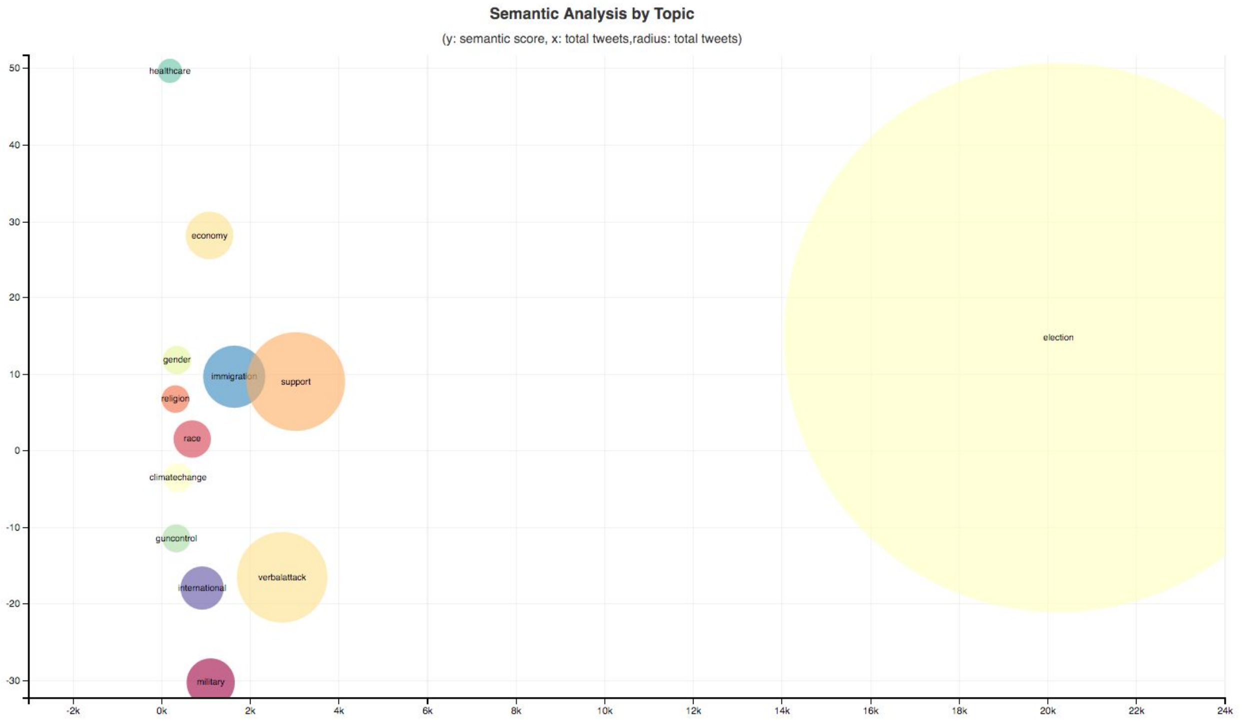


As shown in Graph 5, the radius and x-coordinate of each bubble represents the volume of tweets in this topic, and the y-coordinate represents their semantic score. For all tweets, the semantic score of all topics ranked from high to low as such: health care(average score: 42.7), economy(33.6), immigration(22.9), support(22.5), election(17.7), gun control(13.4), climate change(11.9), gender & sexuality(8.2), religion(7.00), race(6.7), international politics(-2.6), military(-17.8) and verbal attack(-17.9). In other words, the overall sentiment of health care discussion, immigration, election and support is quite positive. Gun control and climate change are similar in their semantic score(slightly above 10) as well as in total volume. Religion, gender & sexuality, and race are also closely positioned groups. International politics has a negative sentiment overall, but it is only weakly negative(-2.6). Military and verbal attack, however, scored less than -17 in semantics, indicating strong negativity of sentiment in discussion.

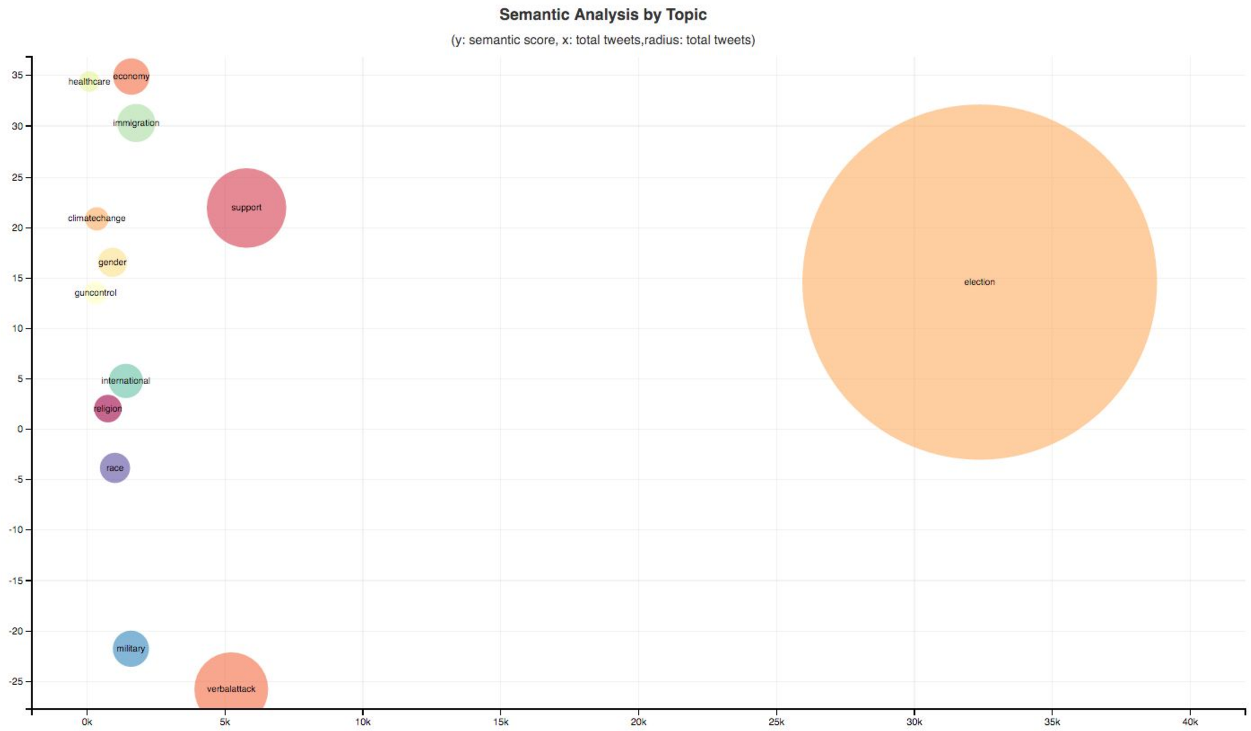
Graph 6: Semantic Score by Topic for Sanders



Graph 7: Semantic Score by Topic for Clinton



Graph 8: Semantic Score by Topic for Trump



Graph 9: Semantic Score by Topic for Cruz

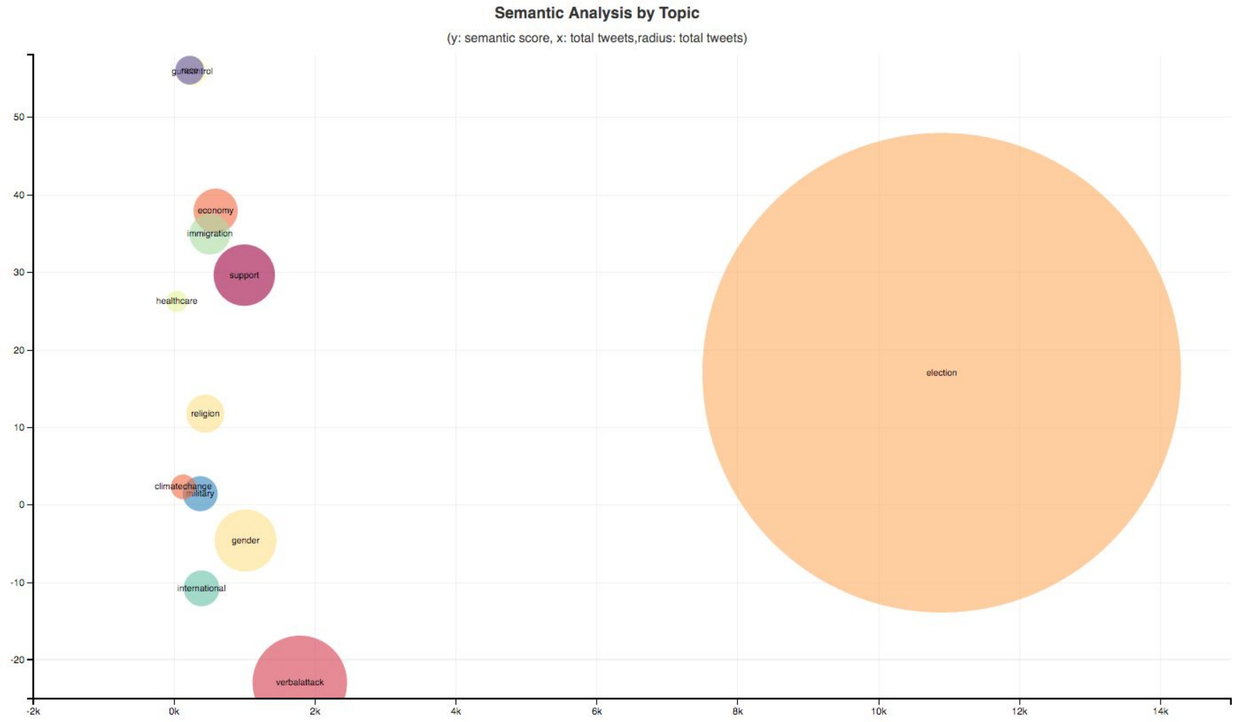
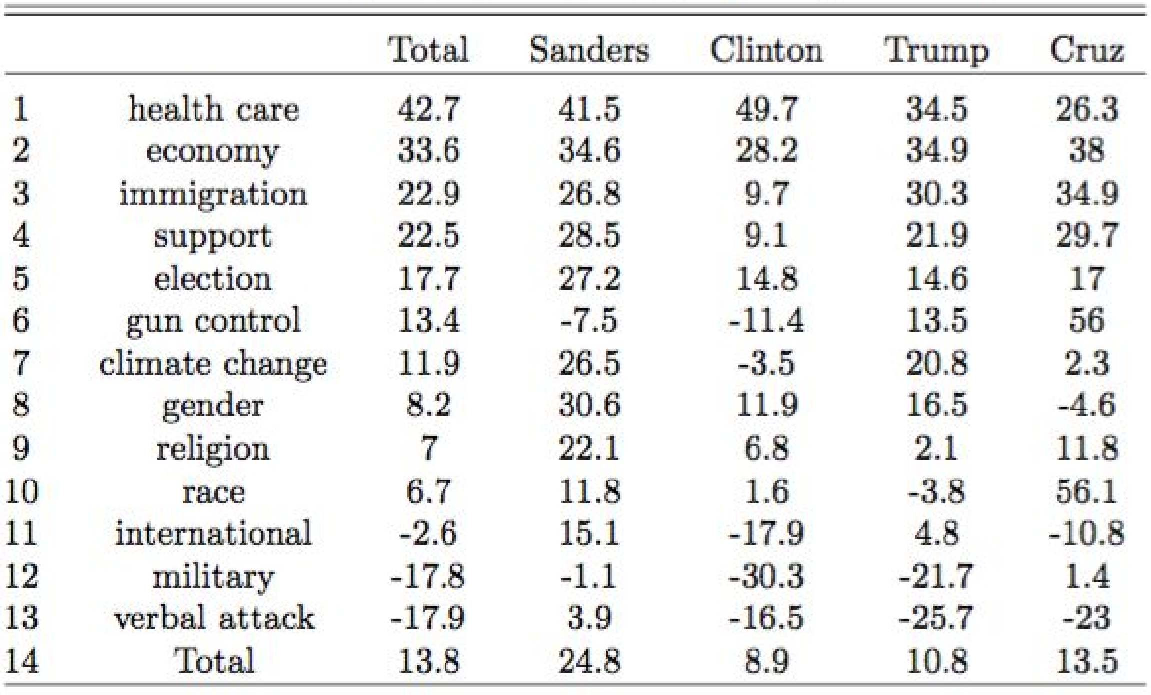


Table 4: Average Semantic Score by Topic by Candidate



		Total	Sanders	Clinton	Trump	Cruz
1	health care	42.7	41.5	49.7	34.5	26.3
2	economy	33.6	34.6	28.2	34.9	38
3	immigration	22.9	26.8	9.7	30.3	34.9
4	support	22.5	28.5	9.1	21.9	29.7
5	election	17.7	27.2	14.8	14.6	17
6	gun control	13.4	-7.5	-11.4	13.5	56
7	climate change	11.9	26.5	-3.5	20.8	2.3
8	gender	8.2	30.6	11.9	16.5	-4.6
9	religion	7	22.1	6.8	2.1	11.8
10	race	6.7	11.8	1.6	-3.8	56.1
11	international	-2.6	15.1	-17.9	4.8	-10.8
12	military	-17.8	-1.1	-30.3	-21.7	1.4
13	verbal attack	-17.9	3.9	-16.5	-25.7	-23
14	Total	13.8	24.8	8.9	10.8	13.5

As illustrated by Graph 6 - 9 and Table 4, The four major candidates are observably different from each other when it comes to semantic score by topic. Overall, people are quite positive when they talk about Sanders on Twitter(24.8), but a lot less positive in discussion about Clinton(8.9). Trump and Cruz have a similar semantic score at 10.8 and 13.t, but they differ a lot in specific topics.

From Graph 6 and Table 4, we found that tweets about Sanders scored notably higher than average in gender and sexuality, climate change, religion, race, international politics and also election. He also has the only positive score(3.9) for verbal attacks. For other topics, he is also higher than average, but only by a smaller margin. The only topic that he scored lower than average is gun control(-7.5 as compared to 13.4). It's worth noting that he has a higher-than-average in the support topic(28.5 as compared to 22.5), which is only matched by Cruz(29.7), which is consistent with his unmatched popularity in the younger generation, the most vocal group represented on Twitter.

Illustrated by Graph 7 and Table 4, Clinton, his major competitor, has the highest score for health care and a close-to-average score for verbal attack, religion, gender and sexuality, economy, and election. However, Twitter discussions which mentioned Clinton by name in topics including military, international politics, climate change, gun control, support, election, economy and immigration all scored the lowest semantic score in all four candidates, which attributed to her lowest overall score. The scale of negativity as well as the wide range of negative topics for Clinton is larger than our expectation, but it could be explained by her former role in the Obama administration and the leaked emails, both of which made the Twitter population doubt her integrity as a politician. We also need to keep in mind that having a negative semantic score does not equal to being viewed negatively in a specific topic for the candidate. It could be that Twitter users are expressing a negative feeling about that specific topic, and mentioned the candidate by name as the possible solver of existing issues.

As for Republicans, Trump has scored surprisingly close to average in topics such as race, gun control, economy and support. Also, he has an unexpected high score in immigration(30.3 as compared to the average of 22.9), climate change(11.9 as compared to 20.8), gender and sexuality(16.5 as compared to 8.2), and also international politics(4.8

as compared to -2.6). While it is far from enough to draw conclusion from these statistics, these unexpected high scores in semantics suggest that people in fact talked quite favorably and positively when discussing Trump's opinion on these topics. Health care, religion, race and verbal attack are topics in which Trump does not do well. Among the four candidates, he is the only one who received a negative score for race, which is as expected considering his controversial remarks on race.

For Cruz, he scored surprisingly high in gun control, support, race and military. In all four topics, discussions about him are all regarded as the most positive of all candidates. While race, gun control and military could be attributed to his relatively favorable view and the speech he gave about veterans and Muslim neighborhoods, the unusually high score in support(29.7 as compared the average 22.5) is hard to explain, and may require further inspection. On the other hand, he has a below-than-average score in topics including health care, climate change, gender and sexuality, international politics and verbal attack. The score for gender and sexuality, as discussed before, is affected by the hashtag #cruzsexscandal, which understandably contains a substantial amount of negative sentiment.

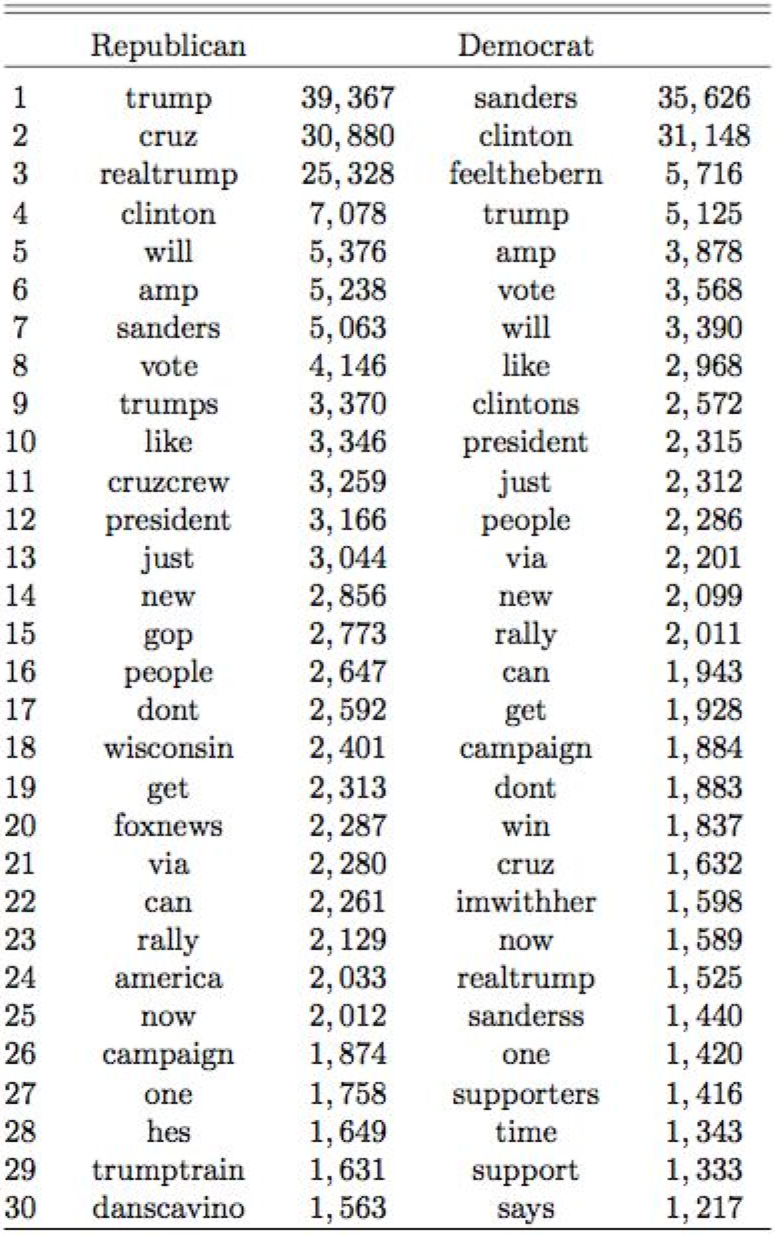
Comparison between Candidates and Parties

Now that we have a better understanding of what Twitter users are discussing when they talk about major election candidates on Twitter, and also how they are feeling in these discussions, we feel it is also necessary to examine to what extent these Twitter discussion are similar or different from each other across candidates. Therefore, we analyzed the most frequent terms found in text for each candidate, and also performed several test of similarity between candidates and parties.

Frequent Term List

We started from generating a list of most frequent terms in tweets for each candidate and each party, and compared them with each other. As shown by Table 5 below, tweets that discussed Democrat and Republican candidate are not fundamentally different from each other in their frequent terms. While for obvious reasons the Democrat group mentioned terms related to Sanders and Clinton quite frequently and Republican group did the same for their candidates, there is an observable overlap of terms related to election. Tweets for both parties used the election vocabulary often, such as “president”, “vote”, and “campaign”. Interestingly, while “trump” ranked the 4th for Democrat tweets, “cruz”only ranked the 21st. For Republicans, the difference between two Democrat candidates are much closer, with Clinton ranking the 4th and Sanders the 7th. Also, “gop” made it to the list for Republicans, but the name of Democrat party is not among the most frequently mentioned terms for the Democrat group. Instead, the Democrats have party neutral terms such as “supporters” and “support” on the list, which do not appear on the Republican list. Another worth mentioning fact is that “foxnews”, the conservative news corporation, is on the list of frequent terms for Republicans, but Democrats do not have a news corporation on the list.

Table 5: Frequent Term List by Party



	Republican		Democrat	
1	trump	39,367	sanders	35,626
2	cruz	30,880	clinton	31,148
3	realtrump	25,328	feelthebern	5,716
4	clinton	7,078	trump	5,125
5	will	5,376	amp	3,878
6	amp	5,238	vote	3,568
7	sanders	5,063	will	3,390
8	vote	4,146	like	2,968
9	trumps	3,370	clintons	2,572
10	like	3,346	president	2,315
11	cruzcrew	3,259	just	2,312
12	president	3,166	people	2,286
13	just	3,044	via	2,201
14	new	2,856	new	2,099
15	gop	2,773	rally	2,011
16	people	2,647	can	1,943
17	dont	2,592	get	1,928
18	wisconsin	2,401	campaign	1,884
19	get	2,313	dont	1,883
20	foxnews	2,287	win	1,837
21	via	2,280	cruz	1,632
22	can	2,261	imwithher	1,598
23	rally	2,129	now	1,589
24	america	2,033	realtrump	1,525
25	now	2,012	sanderss	1,440
26	campaign	1,874	one	1,420
27	one	1,758	supporters	1,416
28	hes	1,649	time	1,343
29	trumptrain	1,631	support	1,333
30	danscavino	1,563	says	1,217

Table 6 presents a frequent term list by candidate, which reveals even more interesting differences between candidates. Trump is mentioned quite frequently when people talked about Democratic candidates(6th for Sanders, 3rd for Trump), but Cruz is nowhere close on the list(not in top 30 for Sanders, and 12th for Clinton), indicating their different level of threat to Democratic candidates. Similarly, Clinton is regarded “People” appeared as the 9th most frequent word for Sanders, but only ranked the 18th for Clinton, 16th for Trump, and 19th for Cruz. “President” ranked the 8th in the frequent term list of Clinton, 9th for Trump, but neither Sanders(the 30th) and Cruz(the 16th) mentions it often. “Feelthebern”, the Sanders’ supporters’ hashtag, made it to the 10th for Clinton, which indicates that Sanders supporters tends to talk about Clinton frequently, but not vice versa. “Obama” appears on Clinton’s list, probably as a result of Clinton’s role in the Obama administration. Sanders, on the other hand, has “democratic” on the list. The Twitter user population is divided when it comes to Trump, as the Trump list has support term such as “makeamericagreatagain”(the 28th) as well as hate message “nevertrump”(the 29th), and also sarcastic term “trumptrain”. Unfortunate for Cruz, his supporter hashtag “choosecruz” only ranked the 24th, but “cruzsexscandal”, a negative message, ranked the 10th.

In short, Twitter users all discussed election news and behaviors(“vote”, “rally”) very frequently when they talked about major candidates, but they also differ in frequent words usage pattern by candidate.

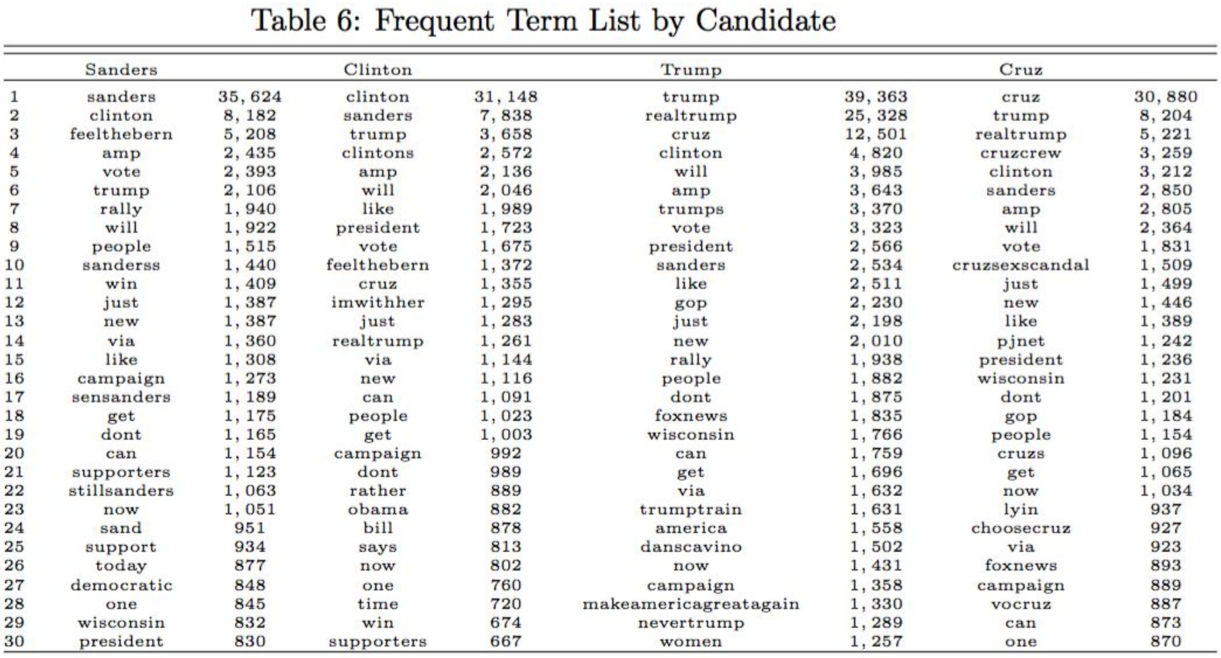


Table 6: Frequent Term List by Candidate

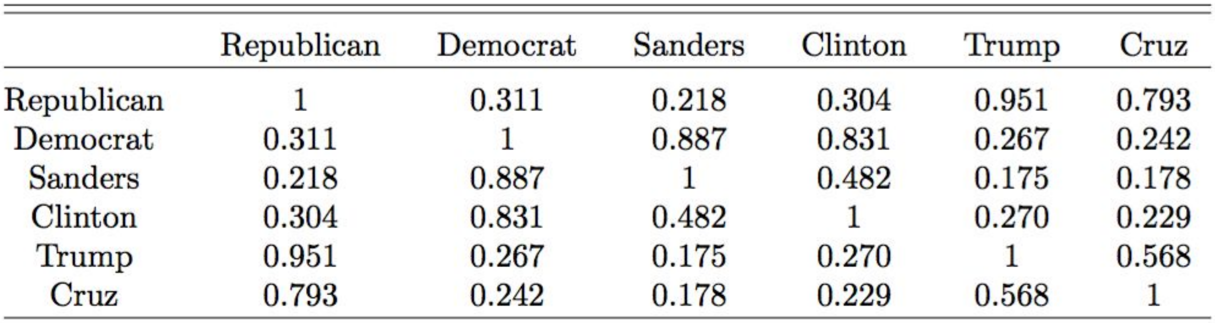
	Sanders		Clinton		Trump		Cruz	
1	sanders	35, 624	clinton	31, 148	trump	39, 363	cruz	30, 880
2	clinton	8, 182	sanders	7, 838	realtrump	25, 328	trump	8, 204
3	feelthebern	5, 208	trump	3, 658	cruz	12, 501	realtrump	5, 221
4	amp	2, 435	clintons	2, 572	clinton	4, 820	cruzcrew	3, 259
5	vote	2, 393	amp	2, 136	will	3, 985	clinton	3, 212
6	trump	2, 106	will	2, 046	amp	3, 643	sanders	2, 850
7	rally	1, 940	like	1, 989	trumps	3, 370	amp	2, 805
8	will	1, 922	president	1, 723	vote	3, 323	will	2, 364
9	people	1, 515	vote	1, 675	president	2, 566	vote	1, 831
10	sanderss	1, 440	feelthebern	1, 372	sanders	2, 534	cruzsexscandal	1, 509
11	win	1, 409	cruz	1, 355	like	2, 511	just	1, 499
12	just	1, 387	imwithher	1, 295	gop	2, 230	new	1, 446
13	new	1, 387	just	1, 283	just	2, 198	like	1, 389
14	via	1, 360	realtrump	1, 261	new	2, 010	pjnet	1, 242
15	like	1, 308	via	1, 144	rally	1, 938	president	1, 236
16	campaign	1, 273	new	1, 116	people	1, 882	wisconsin	1, 231
17	sensanders	1, 189	can	1, 091	dont	1, 875	dont	1, 201
18	get	1, 175	people	1, 023	foxnews	1, 835	gop	1, 184
19	dont	1, 165	get	1, 003	wisconsin	1, 766	people	1, 154
20	can	1, 154	campaign	992	can	1, 759	cruzs	1, 096
21	supporters	1, 123	dont	989	get	1, 696	get	1, 065
22	stillsanders	1, 063	rather	889	via	1, 632	now	1, 034
23	now	1, 051	obama	882	trumptrain	1, 631	lyin	937
24	sand	951	bill	878	america	1, 558	choosecruz	927
25	support	934	says	813	danscavino	1, 502	via	923
26	today	877	now	802	now	1, 431	foxnews	893
27	democratic	848	one	760	campaign	1, 358	campaign	889
28	one	845	time	720	makeamericagreatagain	1, 330	vocruz	887
29	wisconsin	832	win	674	nevertrump	1, 289	can	873
30	president	830	supporters	667	women	1, 257	one	870

Test of Similarity

Though frequent term list is useful for inspection, it is not a quantified examination of similarity between different texts. Therefore, we also performed two tests of similarity on Twitter data: the Pearson correlation test, and the cosine similarity test. The results are shown in Table 7 and Table 8 below.

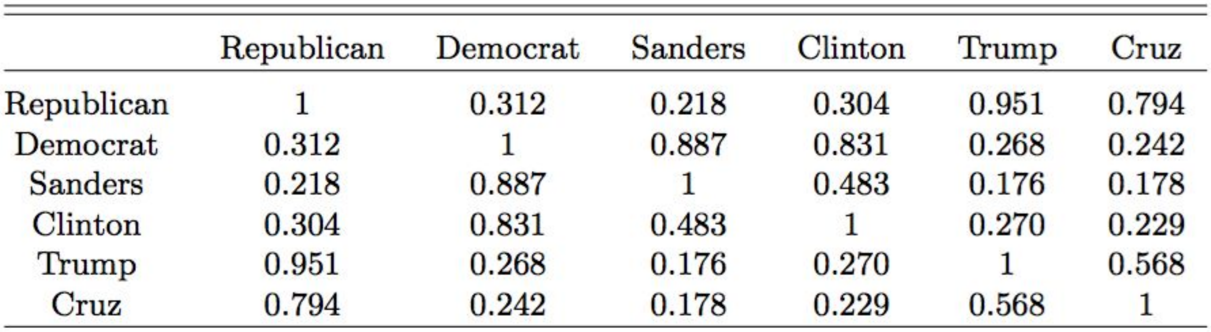
Both tests found that Republican and Democrat tweets are quite different from each other(r = 0.311, cosine similarity = 0.312). For each candidate, Sanders is most different from Trump(r = 0.175, cosine similarity = 0.176), and resembled Clinton the most(r = 0.482, cosine similarity = 0.483). Clinton similarly is closest to Sanders, and lease like curz(r = cosine similarity = 0.229). Trump resembled Cruz the most(r = 0.568 = cosine similarity = 0.568), and Sanders the least. In short, there is a notable difference between two parties, and candidates for each party resembled each other rather than candidates from the other party.

Table 7: Pearson Correlation Test of Similarity



	Republican	Democrat	Sanders	Clinton	Trump	Cruz
Republican	1	0.311	0.218	0.304	0.951	0.793
Democrat	0.311	1	0.887	0.831	0.267	0.242
Sanders	0.218	0.887	1	0.482	0.175	0.178
Clinton	0.304	0.831	0.482	1	0.270	0.229
Trump	0.951	0.267	0.175	0.270	1	0.568
Cruz	0.793	0.242	0.178	0.229	0.568	1

Table 8: Cosine Test of Similarity



	Republican	Democrat	Sanders	Clinton	Trump	Cruz
Republican	1	0.312	0.218	0.304	0.951	0.794
Democrat	0.312	1	0.887	0.831	0.268	0.242
Sanders	0.218	0.887	1	0.483	0.176	0.178
Clinton	0.304	0.831	0.483	1	0.270	0.229
Trump	0.951	0.268	0.176	0.270	1	0.568
Cruz	0.794	0.242	0.178	0.229	0.568	1

Prediction

To qualitatively examine the factors that affect the result of primaries, we use the tweets summary csv files from March 15 to April 15 containing a total of 67,017,623 to develop prediction models. Data collected on primary dates are categorized to testing set. First of all, most relevant predictors are extracted and stored in a data frame. Interactive plots are made to understand to overall trend between predictors and outcomes and hypothesis are developed. Then we apply several algorithms to determine the best-performing variables and construct the optimal models. Initial predictions are conducted with the optimal models. Finally we explore the correlations with first difference approach to account for the change with time trend. One step ahead forecast are also conducted.

Variables

Independent

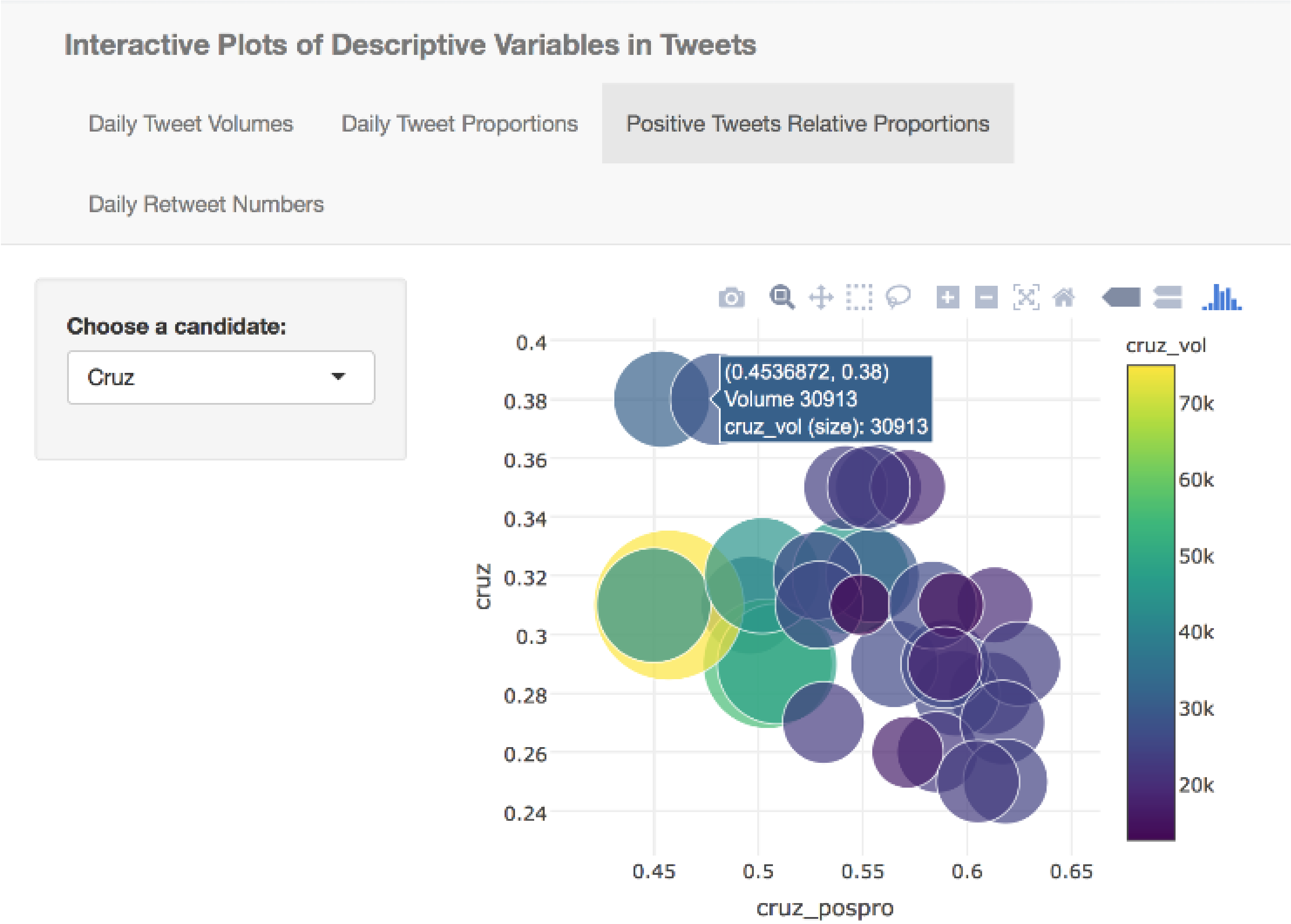
We include variables that examine both magnitude and attitude of the tweets. Pertaining to magnitude, variables are constructed by volumes and proportions. The volume variable counts the daily frequency for each candidate, with the nationwide range. The proportion variable is the ratio of a specific candidate’s tweet volume to the total volume and it is also on a daily basis. To evaluate the twitter user’s attitude, we use a lexicon to identify the words that represent positive or negative emotions in the text part of the tweets. The lexicon contains 6136 words indicating positive and negative polarities and the tweets are categorized accordingly. After obtaining the positive and negative percentages of the tweets, we compute the relative positive proportion by dividing the positive percentage with the sum of both percentages. The original percentages account for only a small portion of all the tweets. Relative proportion gives us a more direct understanding of the overall attitude by comparing positive to negative tweets.

Dependent

Based on daily poll from RealClearPolitics, the approval rate in proportion for each candidate is generated as dependent variable. In addition, the rate is also coded as binary variable such that 1 corresponds to win and 0 corresponds to lose.

To observe the overall trend in poll according to the change in volumes and attitudes, we plot the dependent variable with respect to each of the independent variable. In the interactive plot, the size and color of the marker correspond to tweet volume. Tweets volumes and retweet numbers display similar patterns in plots since increase in retweets leads increase in total volumes. In Clinton's plots, tweet volumes and retweet numbers exhibit similar patterns. The approval rate does not exhibit a clear relationship with the two variables. Highest poll occurs at the medium range in volumes and retweet numbers. The distribution of positive tweets proportion also seems to be random. There is a slight negative correlation with tweet proportion, which implies the discussion of Clinton on Twitter may be negative, resulting in decrease in poll. For Sanders, daily tweet proportion is randomly distributed. The other three variables are somewhat negatively related to poll so that increase in volume, retweets and positively tweets proportion will result in a decrease in poll. In Cruz's plots, positive tweet proportion is also negatively correlated with poll whereas volume, daily tweet proportion and retweets change in the same direction with approval poll. Trump has the most discussion on Twitter as always and the pattern is more difficult to detect. There is no clear trend in any of the variable. One noticeable fact in all four graphs is that the lowest poll corresponds with the brighter and larger markers, indicating high level in volume, proportions and retweets.

Graph 10: Interactive Plot of Positive Tweet Relative Prop for Clinton



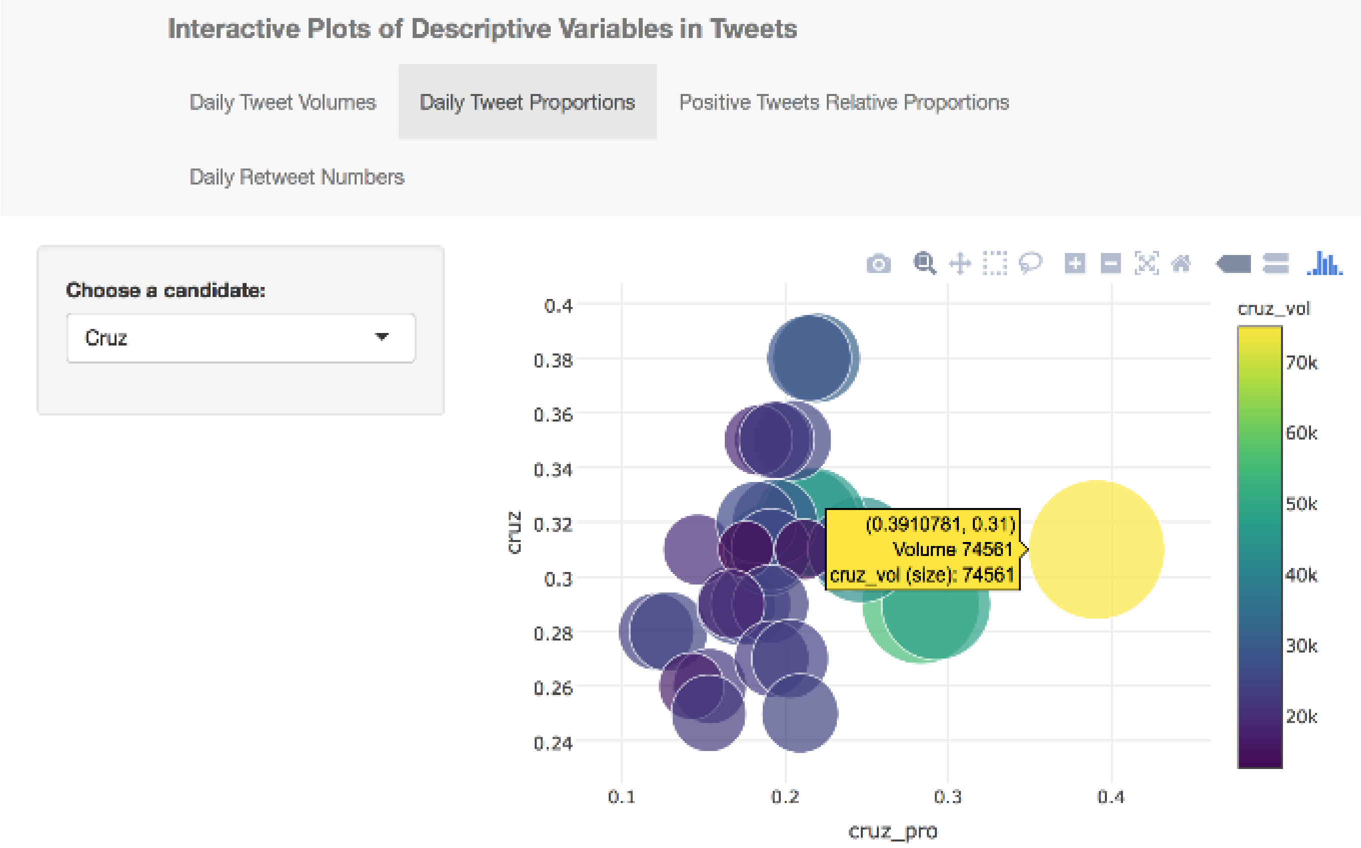
Interactive Plots of Descriptive Variables in Tweets

Choose a candidate: Cruz

[2]

[2] Link for shinyapp: https://xinranwang.shinyapps.io/prediction_plot/

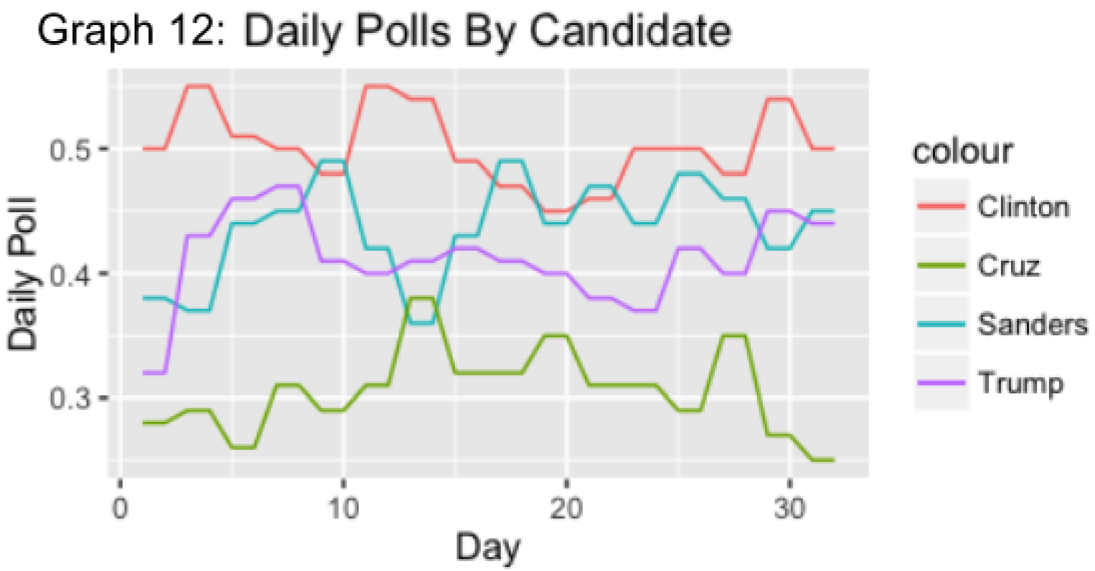
Graph 11: Daily Tweets Proportions for Cruz



Interactive Plots of Descriptive Variables in Tweets

Choose a candidate: Cruz

In the line graph showing the trend of approval rate, each candidate exhibits different patterns. The maximum poll is about 0.55 and minimum is 0.25 with the range of change to be 0.2 at most. For Democrat Party, Clinton has higher polls than Sanders overall. Sanders' poll exceeds Clinton's around March 23 and beginning of April. The most significant change occurs at the end of March when the approval rate for Clinton drops for about 0.15 point and Clinton's increases for about 0.5 point. This change is somewhat unexpected because Sanders won the three caucus held in Alaska, Hawaii and Washington on March 26. The controversy in online poll and actual result suggests that the estimate may fail to capture the real situation. For Republican, Trump is more popular than Cruz throughout the 30-day period that we considered. The difference exaggerates around March 23 and April 15 during which Trump won the Arizona primary on March 22.



Hypothesis

Volume and retweet indicate the popularity of candidates on Twitter but the discussion may not always be positive. Trump, for instance, is the hit of election discussion and is always widely tweeted. But the increase in Trump's tweets may result from the controversial speech that Trump delivered, which very likely stimulates oppositions and reduces his approval rate. Therefore, an increase in volume does not necessarily correspond to an increase in winning poll.

The relationship should be more detectable when specific emotion within the tweets is considered. We expect the poll to be positively related to increase in positive tweets proportion and negatively related to negatively tweets proportion. Relative positive tweets proportion could be ambiguous. The increase in relative positive proportion may indicate increments in both positive and negative tweets with positive tweet increases at a faster rate. However, the increase in overall tweet volume corresponds to more complicated patterns and unpredictable change in outcome polls.

Method

The data set is divided into training and testing sets based on whether primary is held on a given day.

OLS Regression

With actual number as outcome variable, an ordinary least square model is fitted for each candidate. The initial simple OLS regression uses all available predictors: daily tweets volume, tweets proportion, retweet numbers, negative proportions, positive proportion and relative positive proportion. The formula is:

$$Poll = \beta_0 + \beta_1 * Volume + \beta_2 * Proportion + \beta_3 * Retweet + \beta_4 * Negative + \beta_5 * Positive$$
$$+ \beta_6 * Relative\ Positive + e$$

Positive proportion, negative proportion and relative positive proportion are statistically significant predictor for Clinton. The poll for winning increases when positive proportion increases and decreases as negative proportion and relative positive proportion increases. The result corresponds with our assumptions. None of the predictors has statistical significance for Sanders and Cruz. As for Trump, volume, total proportion and retweets are highly statistically significant (Table 9). The supporting rate is positively related with retweet number and total proportion but negatively related with volume. It is also expected as Trump has the highest volumes on Twitter and there is mixed attitude in his discussion. In this case, the increase in volume implies a higher portion of query and incredulity and is negatively correlated with approval rate. To conclude, all six predictors exhibit certain level of statistical significance in regressions and the direction of influence also varies.

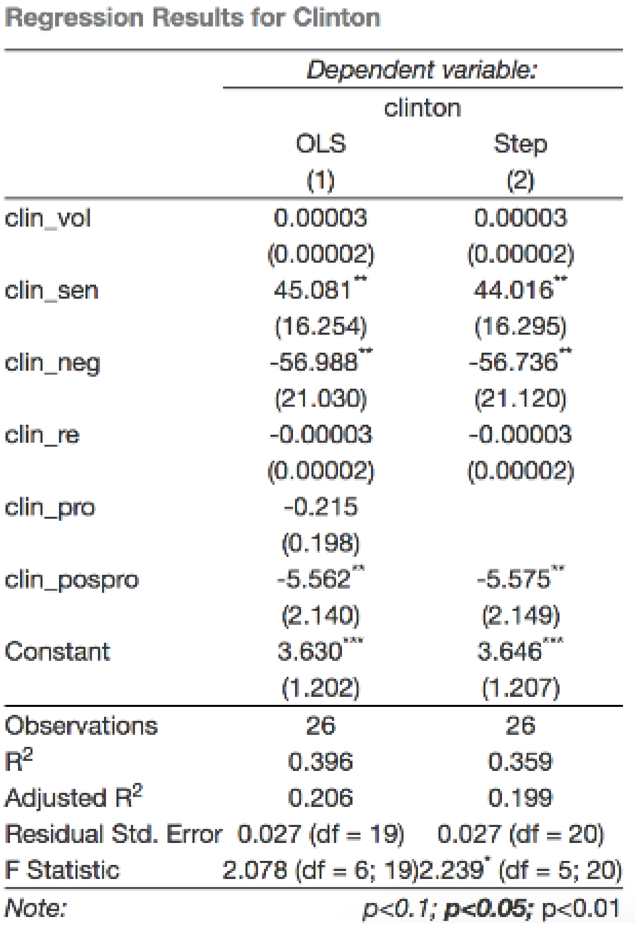
Predictors Selection

Stepwise Algorithm

To sort out the most relevant predictors, we use a stepwise algorithm that combines the forward and backward selection methods and ranks the models by the Akaike

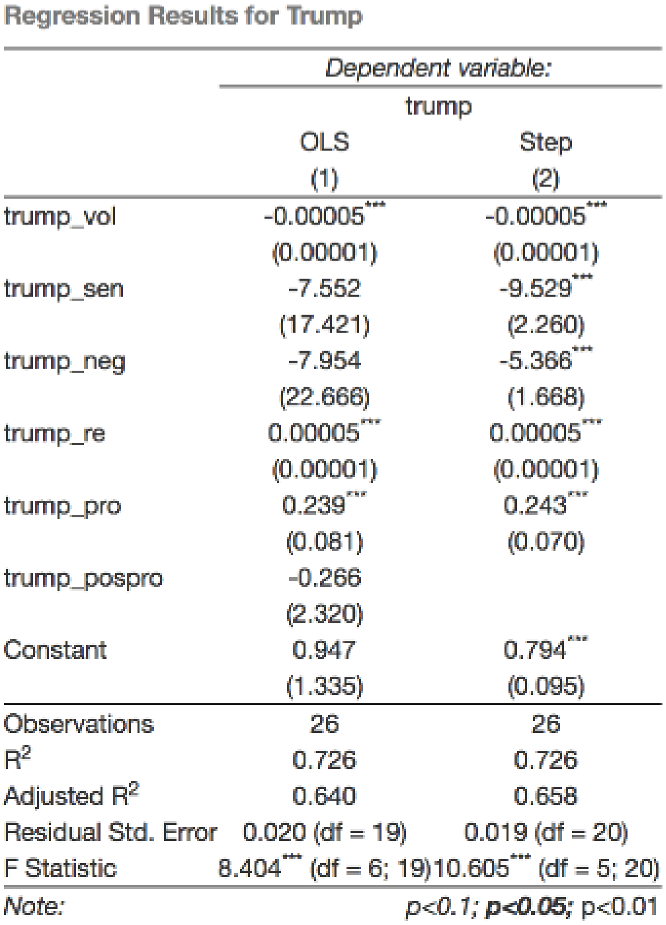
Information Criterion (AIC). The AIC is defined as -2 times the log-likelihood plus 2 times the number of parameters. The linear regression does a better fit in Trump's model since it has more statistically significant predictors and the highest R-squared. We perform the step selection on Trump's model. The stepwise algorithm drops only relative positive proportion and now all predictors are statistically significant. Applying the algorithm to Clinton's model, proportion variable is dropped.

Table 9: Regression Results from OLS and Stepwise Algorithm



Regression Results for Clinton

	Dependent variable:	
	clinton	
	OLS	Step
	(1)	(2)
clin_vol	0.00003	0.00003
	(0.00002)	(0.00002)
clin_sen	45.081**	44.016**
	(16.254)	(16.295)
clin_neg	-56.988**	-56.736**
	(21.030)	(21.120)
clin_re	-0.00003	-0.00003
	(0.00002)	(0.00002)
clin_pro	-0.215	
	(0.198)	
clin_pospro	-5.562**	-5.575**
	(2.140)	(2.149)
Constant	3.630***	3.646***
	(1.202)	(1.207)
Observations	26	26
R^2	0.396	0.359
Adjusted R^2	0.206	0.199
Residual Std. Error	0.027 (df = 19)	0.027 (df = 20)
F Statistic	2.078 (df = 6; 19)	2.239* (df = 5; 20)
Note:		*p<0.1;* ***p<0.05;*** p<0.01



Regression Results for Trump

	Dependent variable:	
	trump	
	OLS	Step
	(1)	(2)
trump_vol	-0.00005***	-0.00005***
	(0.00001)	(0.00001)
trump_sen	-7.552	-9.529***
	(17.421)	(2.260)
trump_neg	-7.954	-5.366***
	(22.666)	(1.668)
trump_re	0.00005***	0.00005***
	(0.00001)	(0.00001)
trump_pro	0.239***	0.243***
	(0.081)	(0.070)
trump_pospro	-0.266	
	(2.320)	
Constant	0.947	0.794***
	(1.335)	(0.095)
Observations	26	26
R^2	0.726	0.726
Adjusted R^2	0.640	0.658
Residual Std. Error	0.020 (df = 19)	0.019 (df = 20)
F Statistic	8.404*** (df = 6; 19)	10.605*** (df = 5; 20)
Note:		*p<0.1;* ***p<0.05;*** p<0.01

LASSO Method

Another way to select the predictors is using least absolute shrinkage and selection operator (LASSO). This method reduces both sum of squared residuals and times the sum of absolute coefficients by performing variable selection and regularization. The formulas are:

$$Poll_{clinton} = \beta_0 + \beta_1 * Volume + \beta_2 * Proportion + \beta_3 * Retweet + \beta_4 * Negative + \beta_5 * Positive + e$$

$$Poll_{trump} = \beta_0 + \beta_1 * Volume + \beta_2 * Relative\,Positive + \beta_3 * Retweet + \beta_4 * Negative + \beta_5 * Positive + e$$

For Clinton, the model drops volume, positive proportion and retweets. We compare the sum of squared error produced by step algorithm and lasso method to determine the model. For both Clinton and Trump, stepwise selections have the smaller sum of squared error so we proceed with the models obtained from stepwise selections.

Prediction

By applying the models to testing data, we obtain the predictions of poll for Clinton and Trump and compare the result with the actual numbers in testing data. For Clinton, the estimated poll corresponds with the actual number overall and the estimated error is 0.05 at maximum. On March 15 and April 9, the predicted polls exceed realclearpolitics poll. On March 15, Florida, Illinois, Missouri, North Carolina and Ohio held primaries and Clinton won in all of them. The predicted poll of 0.55 indicated a strong winning tendency for Clinton. April 9 is the day of Wyoming Democratic Caucus and Bernie Sanders won. It was his seventh straight victory since the Idaho caucus on March 22. Our prediction model fails the capture the real situation. Successive winning may be another influential factor to the result and is a pattern that worth investigating. Correspondingly, time series influence should be considered when constructing the model and this would our next step.

Trump's prediction also exhibits the largest estimated error on March 15. Trump won in four of the primaries held on that day except for Ohio. The predicted approval rate is more than 0.1 higher than the actual poll. During that time, there were four Republican candidates running for president whereas our model only considers the situation with two Republican candidates. The number of candidate affects the outcome and should be included as a control. The over-estimated poll may be due to the lack this information.

First Differences Regression

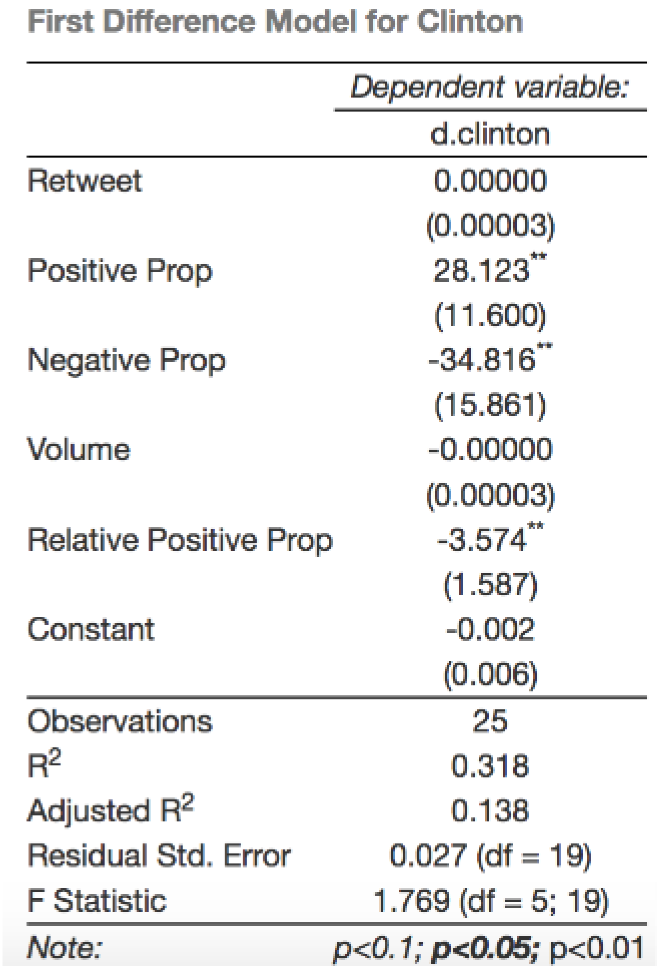
To account for the change with time trend, we conduct a first difference model. For each variable, we subtract the old value at time *t-1* from the new value at time *t* so that:

$$\Delta Poll_{clinton} = \beta_0 + \beta_1 * \Delta Volume + \beta_2 * \Delta Proportion + \beta_3 * \Delta Retweet + \beta_4 * \Delta Negative + \beta_5 * \Delta Positive + \Delta\mu$$

$$\Delta Poll_{trump} = \beta_0 + \beta_1 * \Delta Volume + \beta_2 * \Delta Relative\ Positive + \beta_3 * \Delta Retweet + \beta_4 * \Delta Negative + \beta_5 * \Delta Positive + \Delta\mu$$

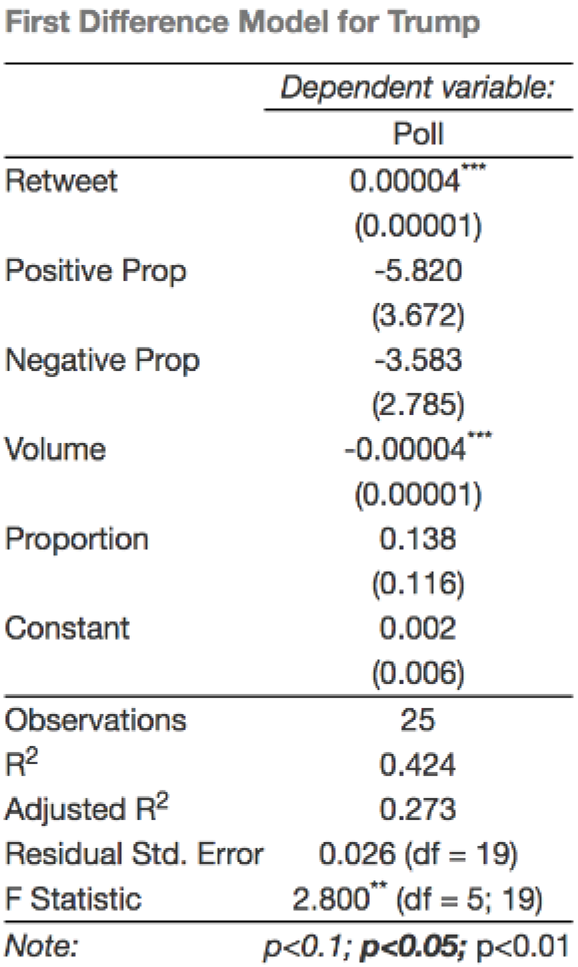
where $\Delta\mu$ is a time-varying error term. We regress the differenced independent variables on the differenced outcome and the effect of changes in poll is estimated on changes in predictors. The first difference model resembles the model developed by stepwise algorithm except for the coefficient on volume. Previous model suggests that increase in tweets volume will lead to increase in approval rate. The first difference model, however, implies that an increase of 100,000 tweets will result in 0.08 point of decrease in poll on average. The first difference model for Trump exhibit similar pattern with the previous model: all coefficients have the same sign and do not vary much in magnitude. But only retweet number and volumes remain statistically significant, which means that the magnitude of tweets is informative when evaluating Trump's poll.

Table 10: First Difference Regressions Results



First Difference Model for Clinton

	Dependent variable:
	d.clinton
Retweet	0.00000
	(0.00003)
Positive Prop	28.123^{**}
	(11.600)
Negative Prop	-34.816^{**}
	(15.861)
Volume	-0.00000
	(0.00003)
Relative Positive Prop	-3.574^{**}
	(1.587)
Constant	-0.002
	(0.006)
Observations	25
R^2	0.318
Adjusted R^2	0.138
Residual Std. Error	0.027 (df = 19)
F Statistic	1.769 (df = 5; 19)
Note:	*p<0.1;* ***p<0.05;*** p<0.01



First Difference Model for Trump

	Dependent variable:
	Poll
Retweet	0.00004^{***}
	(0.00001)
Positive Prop	-5.820
	(3.672)
Negative Prop	-3.583
	(2.785)
Volume	-0.00004^{***}
	(0.00001)
Proportion	0.138
	(0.116)
Constant	0.002
	(0.006)
Observations	25
R^2	0.424
Adjusted R^2	0.273
Residual Std. Error	0.026 (df = 19)
F Statistic	2.800^{**} (df = 5; 19)
Note:	*p<0.1;* ***p<0.05;*** p<0.01

One Step Ahead Forecast

Finally, we build a model to predict the outcome based on lagged values of poll and other predictors. Specifically, autoregressive integrated moving average (ARIMA) model is used. The formulas used are:

$$For\ Clinton:$$

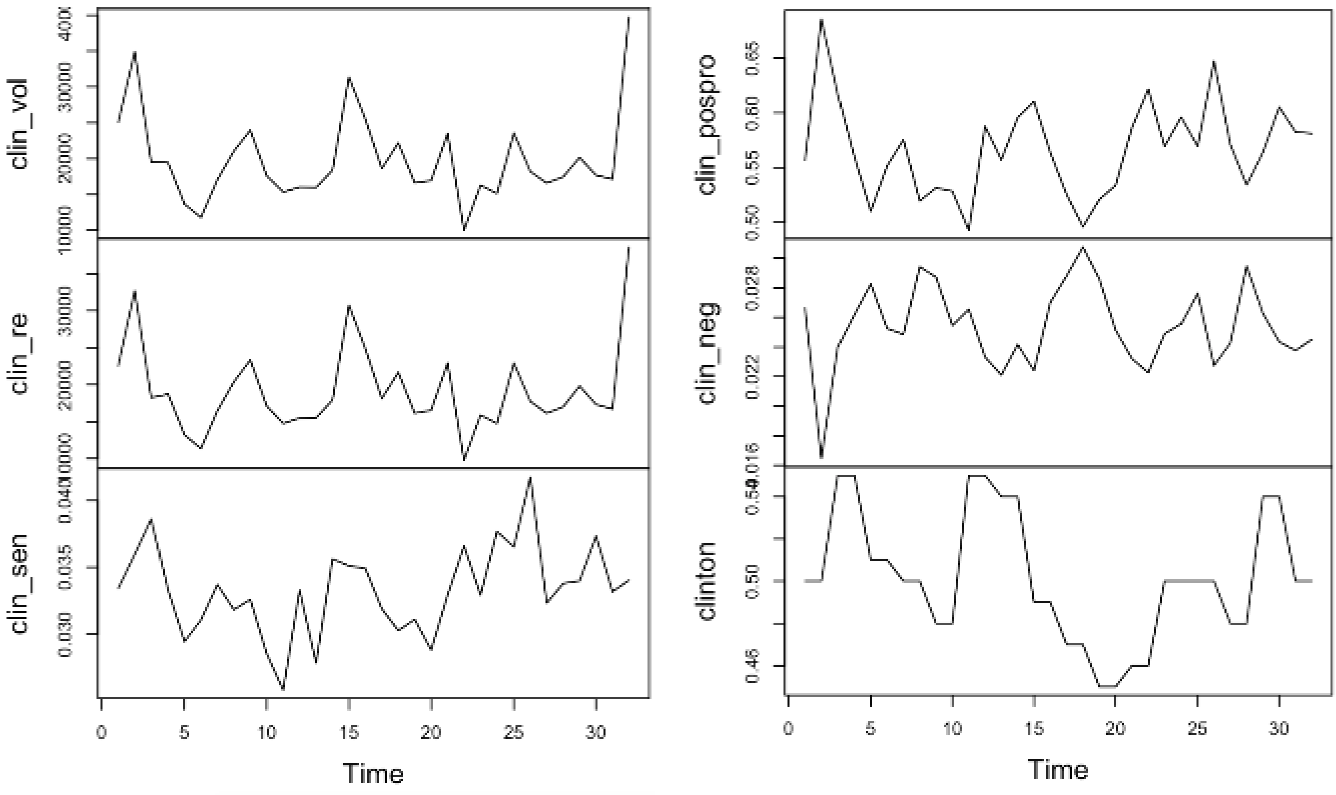
$$Y_t = \beta_0 * Y_{t-1} + \beta_1 * Volume_{t-1} + \beta_2 * Proportion_{t-1} + \beta_3 * Retweet_{t-1} + \beta_4 * Negative_{t-1}$$
$$+ \beta_5 * Positive_{t-1} + e_t$$

$For\ Trump:$

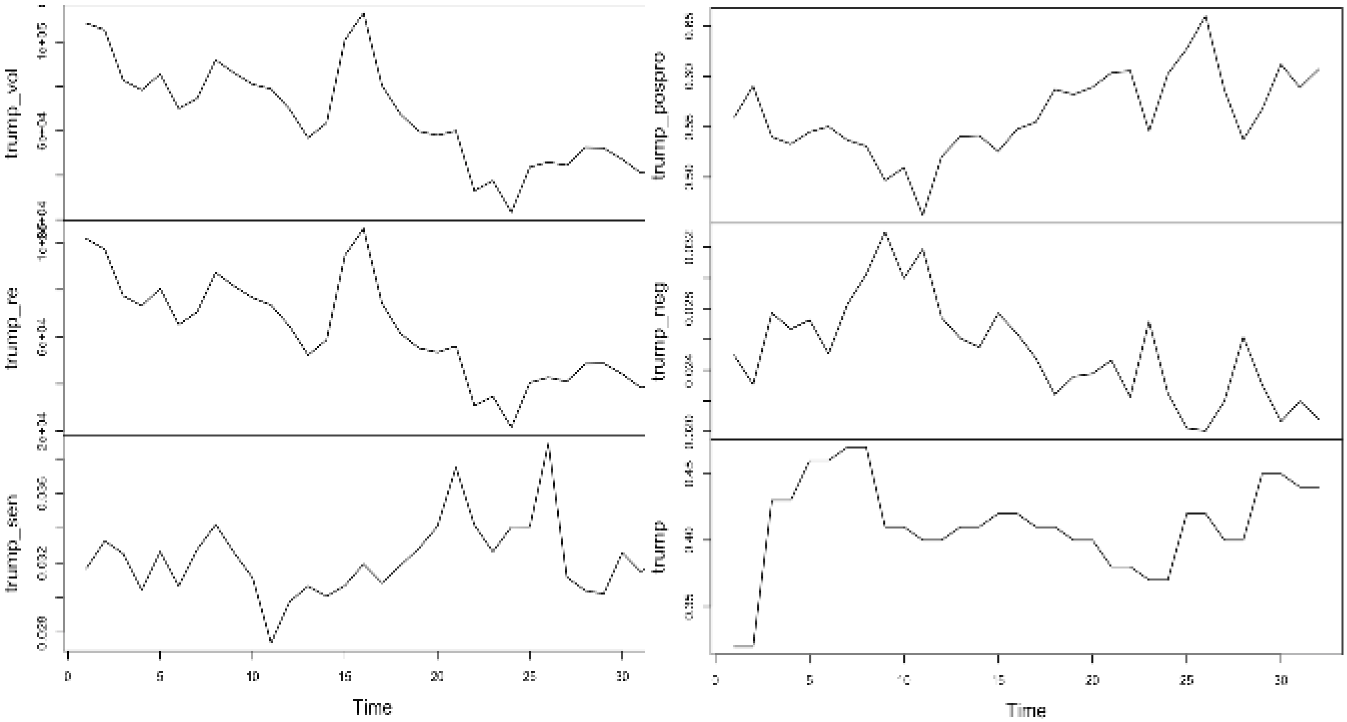
$$Y_t = \beta_0 * Y_{t-1} + \beta_1 * Volume_{t-1} + \beta_2 * Relative\ Positive_{t-1} + \beta_3 * Retweet_{t-1} + \beta_4 * Negative_{t-1}$$
$$+ \beta_5 * Positive_{t-1} + e_t$$

Except for prediction, 95% confidence intervals are also computed and are shown as the black dotted lines in the plot. For both Trump and Clinton, the actual data lies almost completely within the range of confidence interval, indicating that the one step ahead forecast produces credible predictions. Since the lagged variables used are collected just one step ahead and the daily poll for candidate depends heavily on the result of the previous day, the adequate prediction is not unexpected.

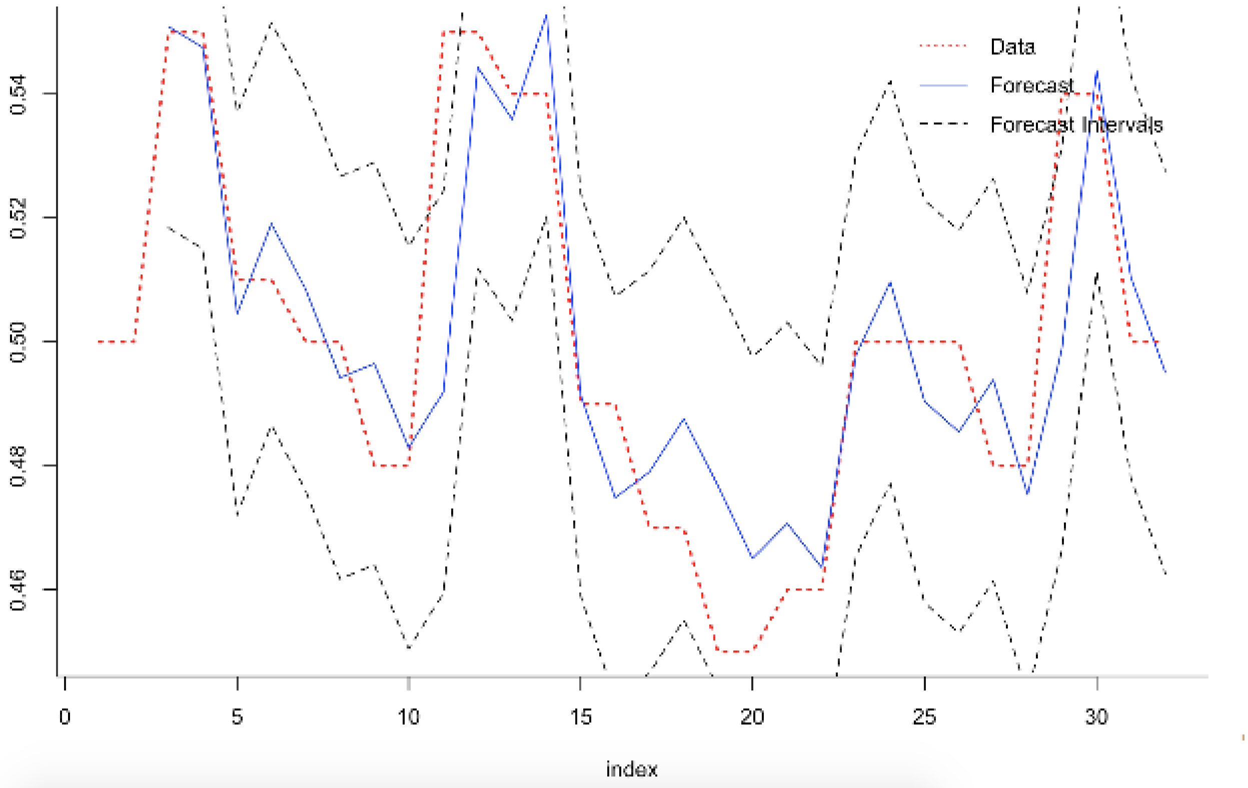
Graph 13: Variables Plot for Clinton



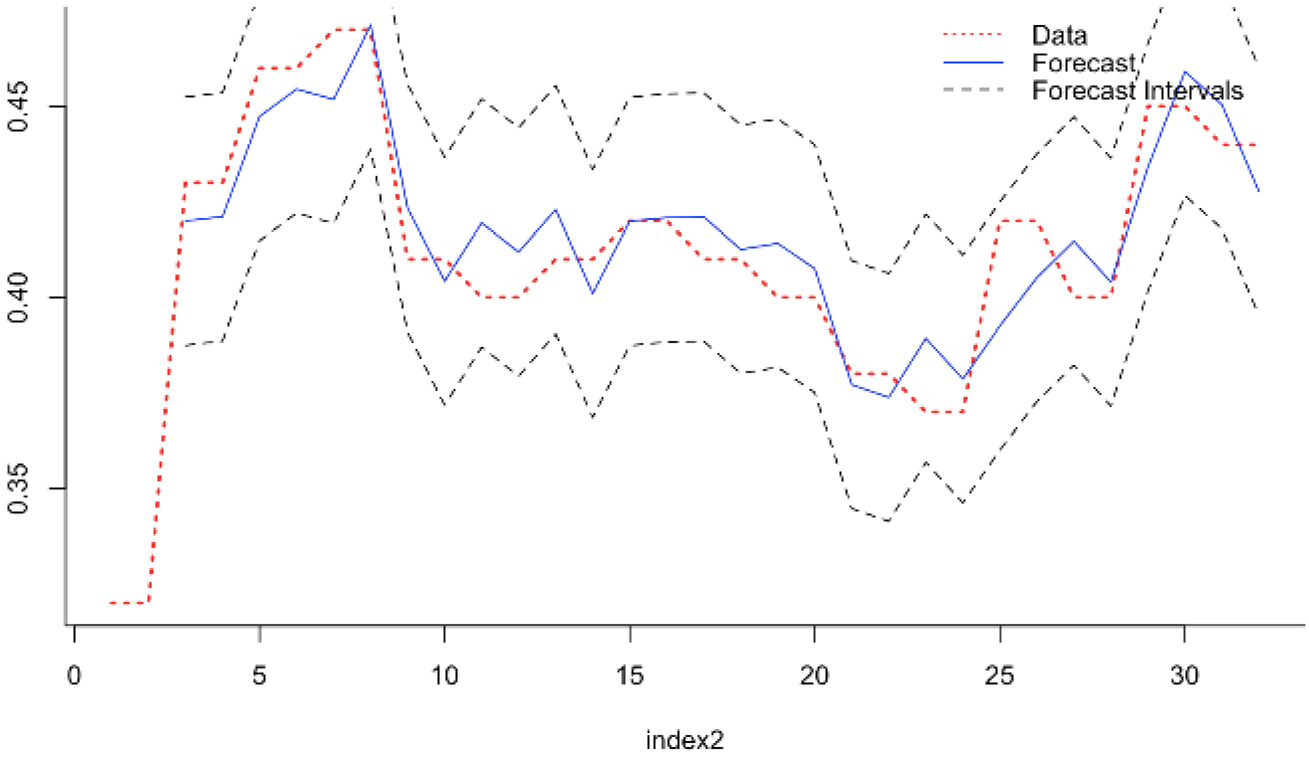
Graph 14: Variables Plots for Trump



Graph 15: One step ahead forecasts for Clinton



Graph 16: One step ahead forecasts for Trump



Conclusion

In this paper, we have demonstrated different approaches to analyze text data and examined the trend in discussions on Twitter regarding 2016 U.S election. Prediction models are also proposed to qualify the trends more accurately. While Twitter data contains demographic bias and may be not to capture public opinion, the result shows a strong correlation between tweet and election result. It implies that the money and time-consuming polling can be supplemented with the analysis of social network and campaign strategy aiming at social media may be further developed.