FINAL PROJECT REPORT

QMSS G4063 Data Visualization

Saad Khalid Alexandra Plassaras Adnan Hajizada Surabhi Bajpai

05/09/2016

The document, represents a final project that aims to analyze tweets centered on the US Presidential Primaries Elections 2016. Our analysis aims to analyze tweets during the election season and link it with associated worldwide events.

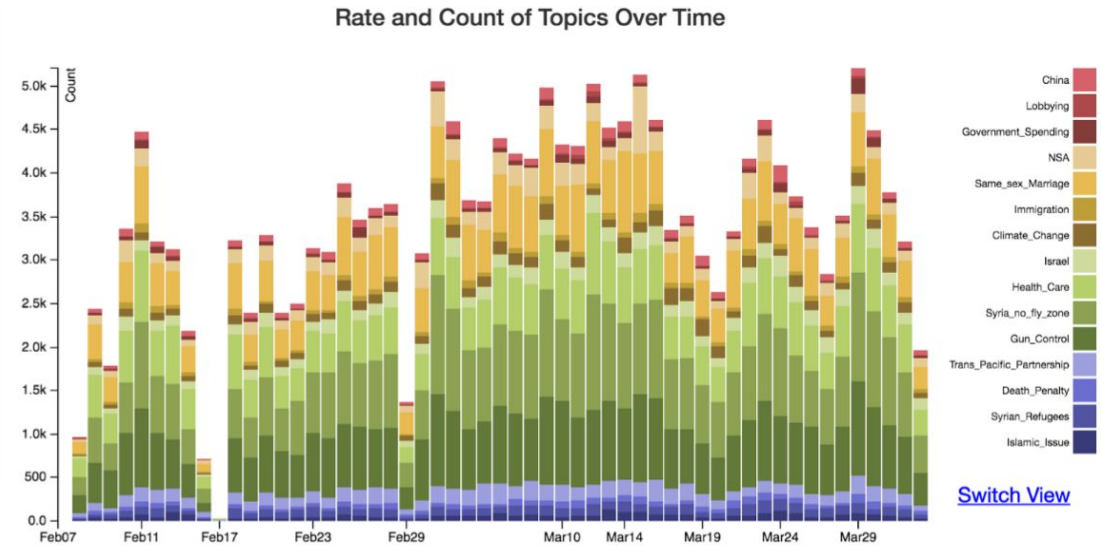
INDEX

Whether or not *real world events* affect Twitter ?

ABSTRACT

The purpose of this paper is to explore Twitter as a way to analyze patterns for the US Presidential Primaries and Caucuses in 2016. Given that this race to the election has seen a pattern of mixed campaign strategy, our paper aims to analyse Twitter feeds from the period starting February 7 to April 2, 2016. Given the scepticism, speculation and engagement attached with the US elections worldwide, the election 2016 has generated a greater sentiment and opinion across Twitter feeds. Moreover, with 200 billion tweets[1] per year across the world, the American population, has been active and vibrant, using social media to express opinions and raise issues with respect to Presidential race for the 2016 election.

Additionally, given that 65%[2] of the American population is active and persistent on using social media, particularly Twitter, using Twitter data to predict elections results has become an increasingly widely used trend, particularly for countries with a large population of active users on social media. In our report below we will attempt to investigate whether real world events are represented in Tweets. To begin, below is a visualization of all tweets collected within our time range and labeled according to the topics that we have created (see issues analysis section below). All visualization shown in this report can be accessed via this link which has various D3 visualizations as well as links to our Shiny app and our github code.



Using the data from the period February 7 to April 2, we are looking to find a correlation with respect to volume and engagement of users and its linkage with the world events and perception of the feeds. Further as reflected in the interactive application created based on the volume of tweets, downsizing it for each candidate, we observe a strong relation between the tweet volume and the date of the primaries.

[1] http://www.internetlivestats.com/Twitter-statistics/#trend
[2] http://www.pewinternet.org/2015/10/08/social-networking-usage-2005-2015/

Moreover, filtering the geo-tagged tweets, we observe that many Twitter users are engaged towards the US Presidential election debate, also reflect sentiments from Europe, Asia and some from parts of Latin America and Africa. Further with respect to the political annotations associated with the election, issues reflecting towards US foreign policy such as those on Israel, the Syrian refugee crisis, Trans-Pacific partnership, Climate change and China. Each of these issues reflect the US association with global sentiments, and given the candidate's inclination and campaign pitch, have also generated sentiments and opinions among the Twitter users.

The reflections by the users have reflected on issues with one user voicing his opinion on Israel as *@dlisraelnews @redbrasco Do you really think Bernie is pro Israel?*, while the sentiments in Europe, also reflect a certain position on the issue, with a Twitter user from Germany commenting *@Salon @stevesalaita when will #FeelTheBern #HillaryClinton .@SenBobCorker .@SenToomey COMPARE Germany #JOBS #WAGES #UNIONS #economy to USA*
on the Twitter sentiment, reflecting a strong voice and opinion on the Syrian refugee crisis. Even though, we have limited the paper with respect to numerical interpretation, skewing towards quantitative analysis, we have also tried to qualitatively include a few extracts of Twitter users with respect to the sentiments being shared and discussed online.

Furthermore, the American sentiment has been interesting to look at given the global association of American voice on these issues. The US map looks largely negative with 73.8 % Percent of sentiments quoting a negative sentiment with respect to the Presidential primary race for 2016.

INTRODUCTION

Social media is being increasingly used to express and voice opinions by active social media users. From what was earlier described as a platform to engage and reconnect with friends, has now become a global phenomenon and a powerful tool to express and voice opinion about pertinent issues and themes. The power of social media was reflected and visible in the association of the Arab Spring and most recently, social media's utility and function has also been reflected to deal with the ongoing Syrian refugee crisis. Using Twitter trends to gauge election sentiment is fairly new and was reflected in the ongoing debate with respect to recent Canadian elections. Andrew Hutchinson's article cites on "Can Twitter data be used to predict elections?"

> *"With 6,000,000+ election-related Tweets sent over the past two-and-a-half months, Canadians have flocked to Twitter to discuss key issues, follow candidate debates and share opinions as the #elxn42 campaign took shape across the nation. By comparison, there were just over 4,000,000 #cdnpoli tweets during the 12 months of 2014."*

This reflects the ongoing usage of Twitter being used to discuss matters related to electoral politics.

In our current analysis we have reflected on the use of social media by linking it to the world events correlated with the US Presidential primary elections. Additionally, our endeavour has been to include and understand the nature of sentiments and the associated relation with the burning issues this election season. As we see in our analysis based on the extracted tweets, issues pertaining (link goes to live D3 visualization) to the Syrian no-fly zone emerges as the top most debated and discussed topic. Though, this is not to say that issues of domestic importance are not being debated and analysed in the current elections. Gun control, health care and same-sex marriages, are also pertinent issues being debated and highlighted during the election debates.

In our research question, we only attempt to correlate these issues and the corresponding global debate. As we discuss in our interpretation, the issue of same-sex marriage certainly reflects global sentiments. Moreover, if only one had to understand the usage of social-media with respect to same sex marriages, the June 26, 2015 judgement legalizing same-sex marriages in the US, by the US Supreme Court was celebrated, shared and cheered across different parts of the world. The sentiment on social media reflected same sex marriages taking over the internet with hashtags ranging from #SCOTUS Marriage, #Gay marriage and #Lovewins. The decision reflected a social media celebration, with the hashtag being used almost 5.5 million in 24 hours, and even the US President Barack Obama tweeting #Lovewins, celebrating the verdict.

The world has been witnessing one of the most severe humanitarian crisis in the world. Given the polarised emotions attached with respect to the mass displacement of refugees in Syria, the political debates and statements have generated social media debates and discussion. Likewise, the election debate in the US has also not been untouched by this issue and Twitter trends surrounding the debate and discussion on the Syrian particularly the decision to embark on a no-fly zone has been of particular

importance. Internationally, German Chancellor Angela Merkel calls for a Syrian-no-fly zone has generated polarised opinions. The issues has also permeated in the 2016 election debate and discussion with Marco Rubio, stating to risk a war with Russia in order to enforce a no-fly zone in Syria. Additionally, these sentiments were also echoed by other candidates like former Florida governor Jeb Bush, former Hewlett-Packard CEO Carly Fiorina, and Sen. Lindsey Graham. Furthermore, Hillary Clinton in her statement has called for creating a "a no-fly zone and humanitarian corridors to try to stop the carnage on the ground and from the air." On the other hand, Bernie Sanders has opposed the general view and sentiment, citing it could get America "deeply involved in that horrible civil war and lead to a never-ending U.S. entanglement in that region." In our Tweets views we see many conflicting opinions. As observed in our data, close to 40,000 of the tweets being centered around the syria-no-fly zone.

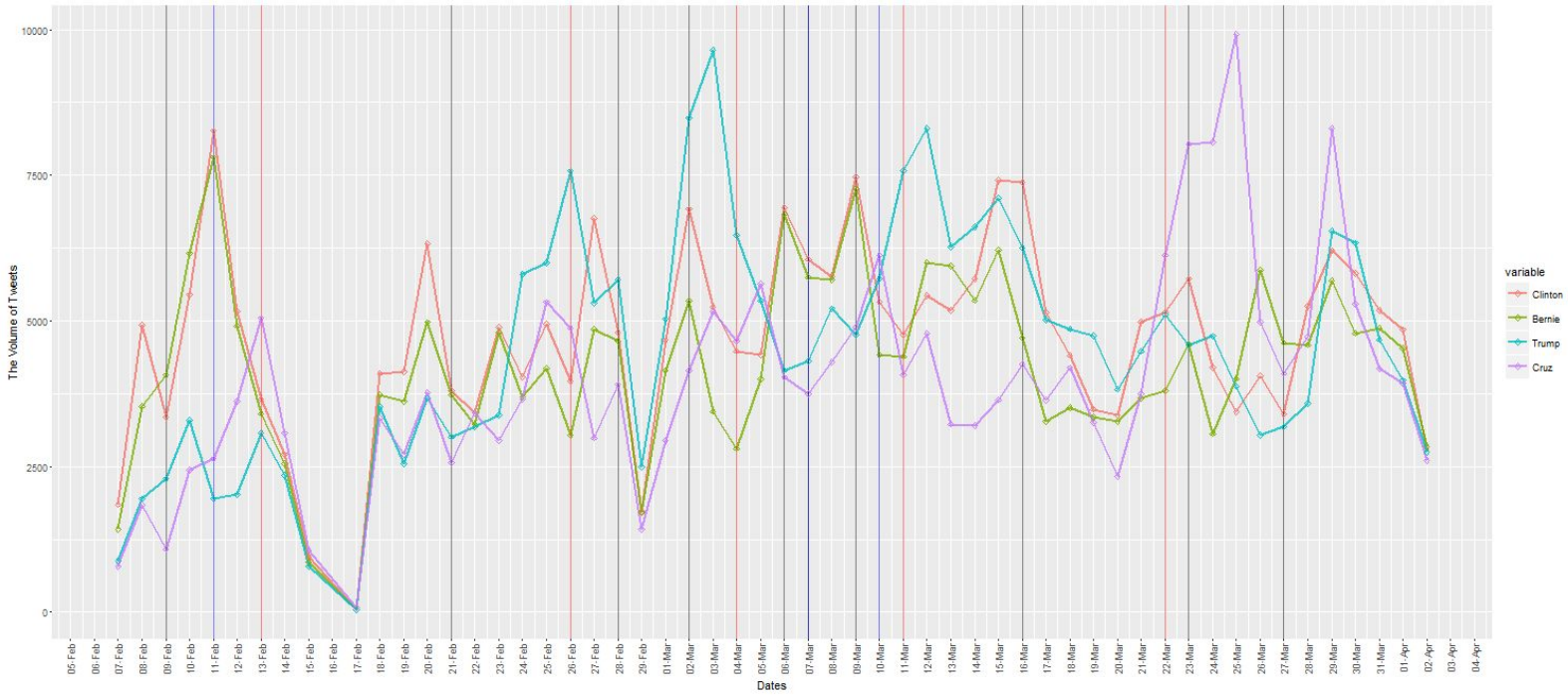
More recently, even domestic scandals have made their way into social media. In this election season, more so from the Republican point of view there have been personal attack by candidates, smear campaigns being run by candidates against each other. Likewise, these events have also attracted great social media attention and have people commenting and tweeting on these campaigns. For example, if one takes into consideration the volume of tweets, Ted Cruz's volume of tweets show a spike on March 25 which coincide with the personal attacks of Ted Cruz blaming Donald Trump for lying. The sentiment of running a smearing campaigns was also mirrored by Donald Trump, accusing Ted Cruz of having an extra-marital affair and launching a personal attack against his family. The approach of drawing skeletons from a candidate's clauset, garnered much public attention with close to 10,000 tweets being driven by Ted Cruz alone.

THEORY: DESCRIPTION OF HYPOTHESIS

Our research question entails, correlating the issues and discussions, of the US 2016 primary elections, with international events across the world. Given the multi-faceted and diverse nature of the US population and a global melting point of people from all over the world, the natural tendency during the presidential election issue is often to cite US-standpoint on issues of international relevance. Moreover, the public opinion on weighing the candidates is often titled on the US standpoint on issues relating to international importance. This seems to have become more important in the post 9/11 world and after the US invasion of Iraq and the ongoing battle in Afghanistan. Moreover, this is also reflected in the echoing sentiments on pro-Israel stand often reflected in the statements and policies of the US.

Social media is increasingly connecting the world in many different ways. The election debates and election issues have realised the importance of social media and therefore, these tools are starting to be used as active election campaign strategy, particularly after the success of the Obama campaign, largely being crowdsourced and people-driven. These sentiments are also reflected in current campaigns, particularly by Bernie Sanders, in his messaging of trying to gain public support. Hence, quoting primary debates, the primary results and their correlation on election issues was of particular interest for our analysis.

Our research and methodology has been sequential. We first analyzed the volume of tweets pertaining to each candidate, filtering the raw data shared for the project. This was primarily done by subsetting the data according to the key-words used and suggested in the first assignment and during the course of this class. Therefore, we have identified our filtered tweets, via the names of the respective candidates, their commonly used Twitter tags and the campaign slogans often quoted and used on Twitter. Given volume of tweets, we collected the respective number of each of the candidates. As a way to normalize this we have considered this to be in percentage terms for the total volume of tweets. Therefore, our final tally is represented by the graph below.

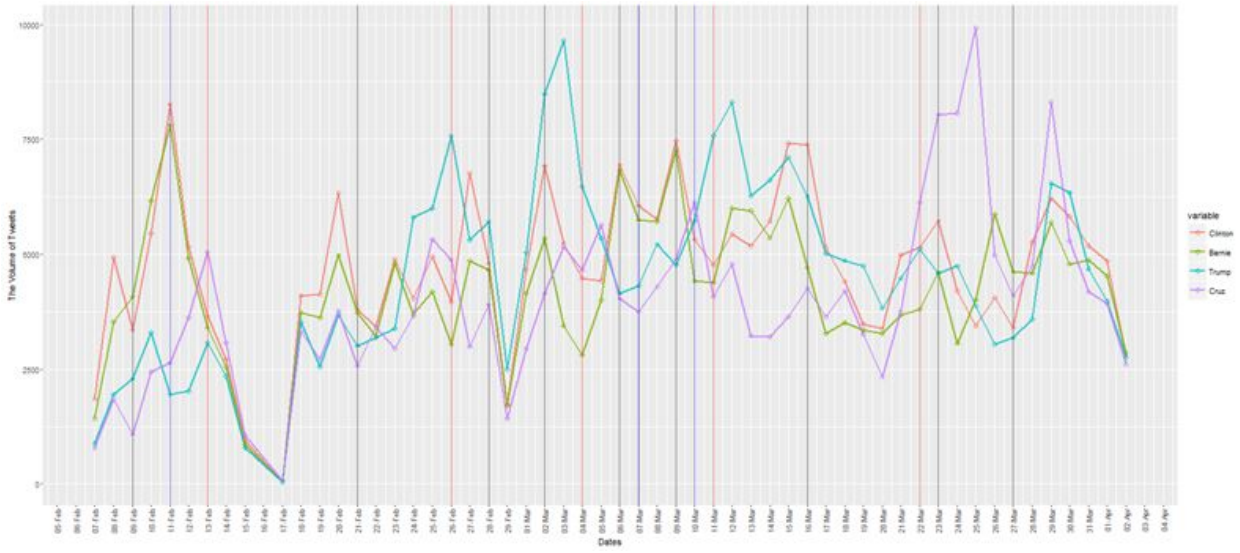


Note: As a methodology, only the given volume of tweets for the last 4 remaining candidates was used. Data was collected and analysed before Ted Cruz announced the suspension of his campaign.

PRESENTATION OF ANALYSIS

VOLUME ANALYSIS

While looking at the plots representing the volume of tweets that mention certain candidates it is important to say that the files that were provided to us did not have any tweets on February 17th. Authors of the paper debated whether there is a need to fill the gap with estimate, random or an average data. At the end the decision was made to keep it empty to preserve the integrity of the analyses. In our line plots each candidates is represented with a different color. On top of that, you can see vertical lines that were superimposed on the graph. These are some real life events. Black lines represent primaries. Red Lines represent Republican debates and blue lines represent Democratic Debates. One of the major themes of our analyses is to look at how real world events affect the volume of tweets and the sentiment of tweets.



Looking at the plot of the tweet volume for all of the candidates we can already discern some initial patterns. Perhaps not surprisingly, we can observe a "local maximums" for Clinton and Sanders on 11th February during the Democratic Debate. The volume of tweets about Donald Trump peak on February 26 and March 3 during (or a day before) the Republican Debates. If we look at the Ted Cruz's volume of tweets we see that there is a spike on March 25th. This is the day when Ted Cruz decided to blame Donald Trump for lying and starting a smear campaign against Cruz for accusing him of extra marital affairs. There was an allegation about this in the National Enquirer and Cruz was suggesting that this campaign was organized by Donald Trump.[3] The story about extra marital affairs and mutual accusations could have spiked the interest of the public hence the volume of discussion about Cruz.
Sample from March 25th:

[3] http://www.bbc.com/news/election-us-2016-35899703

To me the #CruzSexScandal is the notion that any woman would willingly have sex with Ted Cruz.

March 29th also exhibits the peak in tweets about Ted Cruz, and that is the day of the Republican town-hall in Milwaukee, Wisconsin. Trump, Cruz and Kasich all participated in this town-hall that was hosted by CNN's Anderson Cooper.[4] It is important to note that this high interest in Cruz during the town-hall correlated with his eventual win of the Wisconsin primary on 5th April with close to half of the vote and gaining 36 delegates out of 42 (Trump being far second with only 6 delegates). It is important to note here that some social scientists such as Fabio Rojas from Indiana University suggest that merely the volume of the tweets about the candidate can predict her victory on the elections. With the main hypothesis being that, "If people must talk about you, even in negative ways, it is a signal that a candidate is on the verge of victory."[5]

SENTIMENT ANALYSIS

The sentiment analyses was conducted by combining the sentiment lexicon developed by Pablo Barbera and opinion lexicon by Hu and Liu (2004)[6]. According to the list of words that have positive and negative connotation we have calculated the sentiment ratio for each candidate on each date that we were provided the data about. Our sentiment score is calculated as:

(Positive Score – Negative Score) / Total Score.

This equation gives us a ratio of scores between -1 and 1 that is a good predictor of the sentiment from very negative to very positive. The data that we have acquired ended up being predominantly negative. This could be due to a general notion of negativity online that is well described by various authors and researchers.[7] For the benefit of our analyses we have normalized the score around the zero mean and scaled it for the minimum and maximum of -1 and 1. This will give us a better understanding of the mood and change of the attitude towards the candidates.

If we look at the overall sentiment plot of all of the candidates we can clearly discern some patterns. On average the tweets about Ted Cruz are mostly below the mean which means they are predominantly negative. There is a general negative sentiment about Ted Cruz among his co-workers and among the people who knew him personally[8] it looks like this sentiment has transferred well to the world of Twitter. Overall, especially if we look at the second part of the graph starting from first of March we see that

---

4 http://fox6now.com/2016/03/24/cnn-to-host-gop-presidential-town-hall-live-from-milwaukee-on-march-29th/

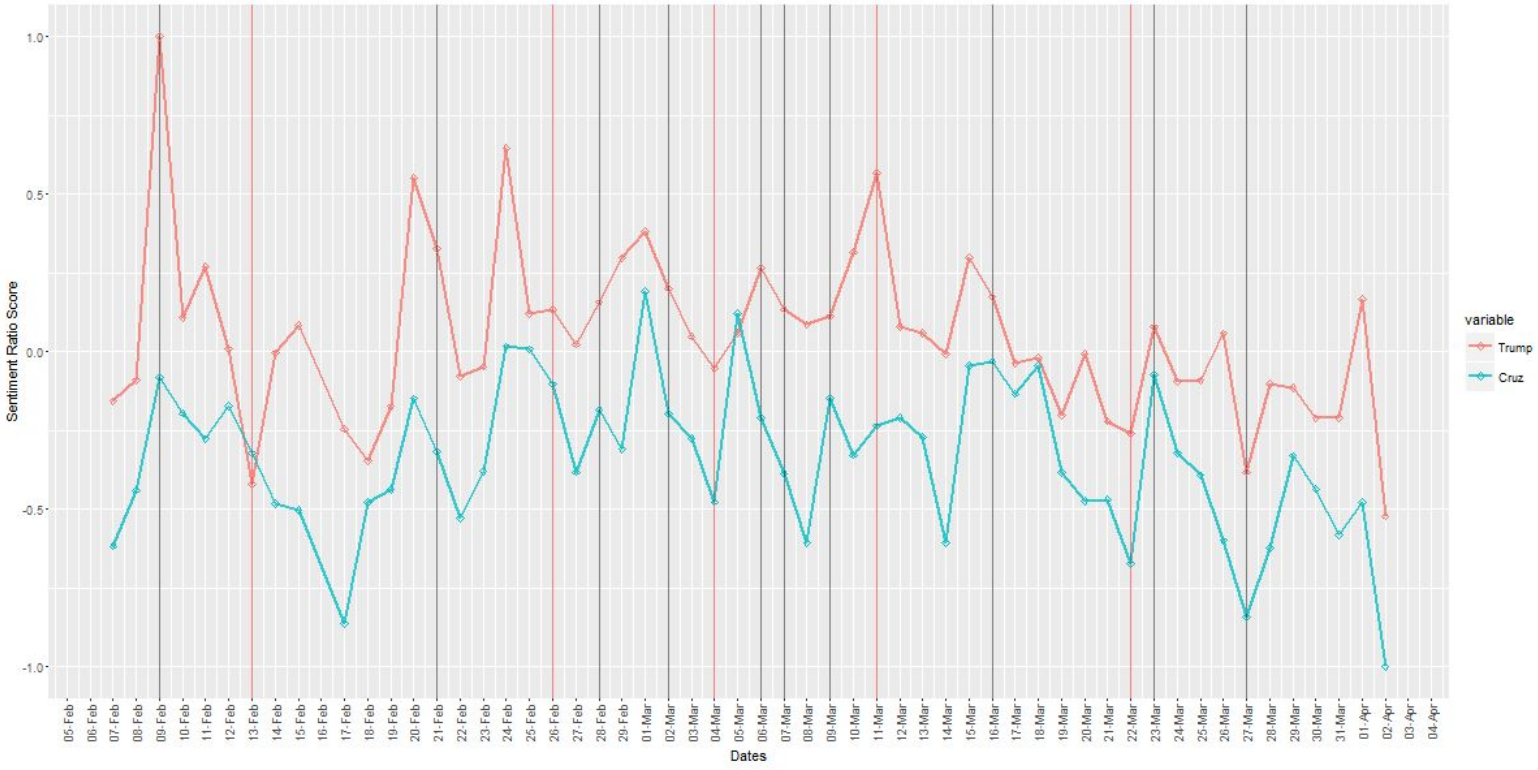
5 https://www.washingtonpost.com/opinions/how-Twitter-can-predict-an-election/2013/08/11/35ef885a-0108-11e3-96a8-d3b921c0924a_story.html

6 https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

7 http://www.bbc.com/news/technology-31749753

8 http://www.telegraph.co.uk/news/worldnews/us-politics/12121499/Why-do-so-many-people-hate-Ted-Cruz.html

Bernie Sanders is the candidate who probably is the most popular and enjoys a plethora of positive tweets about him. By now, it is clear that Bernie Sanders is "winning" among the young voters.[9] Due to that fact that more than a third of Twitter users are young (18-29 years of age), Sanders's "winning" in the real world is visibly transferred to the online world and Twitter particularly.



While examining the Sentiment Ratio Scores of Republican candidates we can clearly see that throughout the data both lines are somewhat correlated as in the increases and decreases in the sentiment happen around the same peaks.These lines represent corresponding trends, with peaks and valleys coinciding on certain dates. This can be an indicator that fluctuations in the sentiment score are closely connected to the attitude to the Republican side as a whole rather than a specific candidates and the difference between two lines just shows how much Donald Trump is more popular than Ted Cruz. In fact, the correlation coefficient between two sets of data is 0.629 which proves our hypothesis. When it comes to real world events, both candidates show no indicators of change during the Democratic Debates thus we did not include those lines into our graph. Tweets about Donald Trump show the highest sentiment score during the New Hampshire primaries on February 09. Trump won those primaries, getting more than 35% of the vote. This was his first win, and that is the connection between the peak in sentiment. Trumps drop in sentiment score coincides with the February 13th Republican Debate. Another peak comes just 1 day before the 21 February primary and Republican Debate of March 11th. Both candidates see a drop in

9 https://www.washingtonpost.com/news/the-fix/wp/2016/03/17/74-year-old-bernie-sanderss-amazing-dominance-among-young-voters-in-1-chart/

sentiment score during the 26-27 March range. Interestingly enough there was only Democratic Primary on those days, so the maybe people who were active online on those days were generally negative about the republican nominees.

When looking at the sentiment plot of the candidates from Democratic Party we decided to include the Republican Primaries because it seems that they also had an effect on the sentiment scores of Clinton and Bernie Sanders. When we examine this plot we can see that the correlation coefficient between these two lines 0.37 which is much lower that with Republicans. We can suggest that on average there is much more variation between sentiment towards candidates from the Democratic Party than between the candidates from the Republican Party. Tweets about Bernie Sanders have a very high sentiment score on the February 9th which is the date of the New Hampshire primary. Bernie Sanders actually won those primaries and gained more than 60% of the vote. Bernie Sanders' sentiment score drops during the Republican Debates of February 13th, 26th and March 4th. This can be explained by the abundance of conservatively minded Twitter users who were watching the debate and were generally negative towards Bernie Sanders even though he was not a participant of the Debate. Interestingly enough, the Democratic debates of the March 7th and March 9th also brought a drop in a sentiment score for both tweets about Hillary Clinton and Bernie Sanders. This could happened because the supporters of each candidate were attacking the other candidate during the debate.

Tweets about Hillary Clinton demonstrated higher sentiment score before the February 20th primaries that happened in Nevada where Hillary Clinton won by gaining more than half of the vote. Her sentiment score drops the lowest before the Republican Debate of 26th February. This is the debate where Republican candidates attacked Hillary Clinton due to her actions during the Benghazi crisis and the

scandal with using work emails on a personal server. We believe that the republican users were attacking Hillary Clinton online feeding off the topics from the debate

ISSUE ANALYSIS

Below is a visualization of all the tweets collected during February 7th to April 2nd. In our analysis we chose to create a lexicon that could better understand what was being tweeted online in regards to our four presidential primary candidates. Thus the following lexicon was made:

Term	Words the term was comprised of
Islamic Issue	ISIS, destroy, aggressive, bomb, deter, contain, lessons, past, alone, jihad, commander, chief, death, warrant, signing
Syrian Refugees	65, refugees, screening, imagine, magic, number, responsibility, helping, weakness, sending, out, persecution, christians, targeted, genocide
Death Penalty	Egregious, consideration, state, democratic, civilized, society, murder, americans, believer, horrendous, crime, punishment, human, life, ultimate
Trans-Pacific Partnership	Agreement, bar, high, continuation, disastrous, trade, policies, cost, millions, bottom, negotiate, strong, conservative, country, better
Gun Control	Many, people, die, middle, ground, guns, tremendous, regulation, opponents, weak, second, ammendment, very, strong, bad
Syria-no fly zone	Advocating, humanitarian, corridors, carnage, stop, unilateral, no-fly, horrible, civil, war, entanglement, sit, back, business, sticking our nose
Health Care	Deductible, rising, out-of-pocket, costs, right, all, people, replace, terrific, referendum, obamacare, flat, tax, intend
Israel	Israel, alarmed, regime, existence, destruction, denies, nation, never, more, jeopardy, seek
Climate Change	Choice, energy, consume, produce, climate, change, threatening, planet, horrendous, clean, air, water, environmental, important, believer
Immigration	Strengthen, families, finally, fix, broken, immigration, system, aggressive, policies, southern border, Mexico, wall, humane, blatantly, laws
Same-sex marriage	Ban, discrimination, L.G.B.T., americans, everybody, sanctity, human, life, uphold, life, sacrament, marriage, learn, live, marry, defend
NSA	Betrayed, surveillance, program, Patriot Act, voted, against, major, reform, conversation, listening, truth, security, terrorists, constitution, law-abiding
Government Spending	Core, issues, rails, against, reckless, government, spending, abolishing, agencies, education, department, internal, revenue, service, appearance
Lobbying	Blamed, Washington, cartel, lobbyists, k street, corruption
China	China, campaign, territorial, expansion, intent, kick, America, out, the, Pacific, on

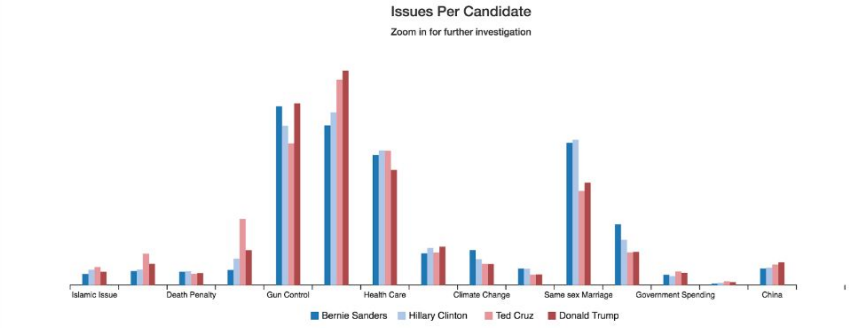
As a way to highlight the analysis of the issues, we initially tried locating the lexicons via an online source. However, the search online did not result any matched requirements. Therefore, as a way to look at multiple issues, we created a lexicon based on the quotes given by the candidates, with respect to the elections. New York Times proved to be a valuable source, since we were able to categorize quotes given by each of the candidates for each issues. Initially, we considered dividing the lexicon according to each candidate and category, but given that tweets are often clustered around issues and concerns raised by different candidates, we combined the different words and quotes, and therefore, combined the different quotes of each candidates and created a combined lexicon. This helped expand the issue index and the vocabulary list and was useful in capturing multiple issued featuring different issues.

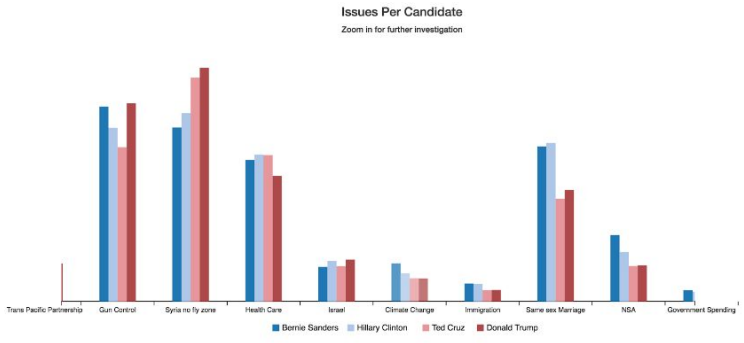
The top topics throughout the whole time range were:

- Syria no fly zone
- Gun control
- Health care
- Same sex marriage

While the topics least talked about were:

- Lobbying
- Immigration
- Death Penalty
- Islamic Issue



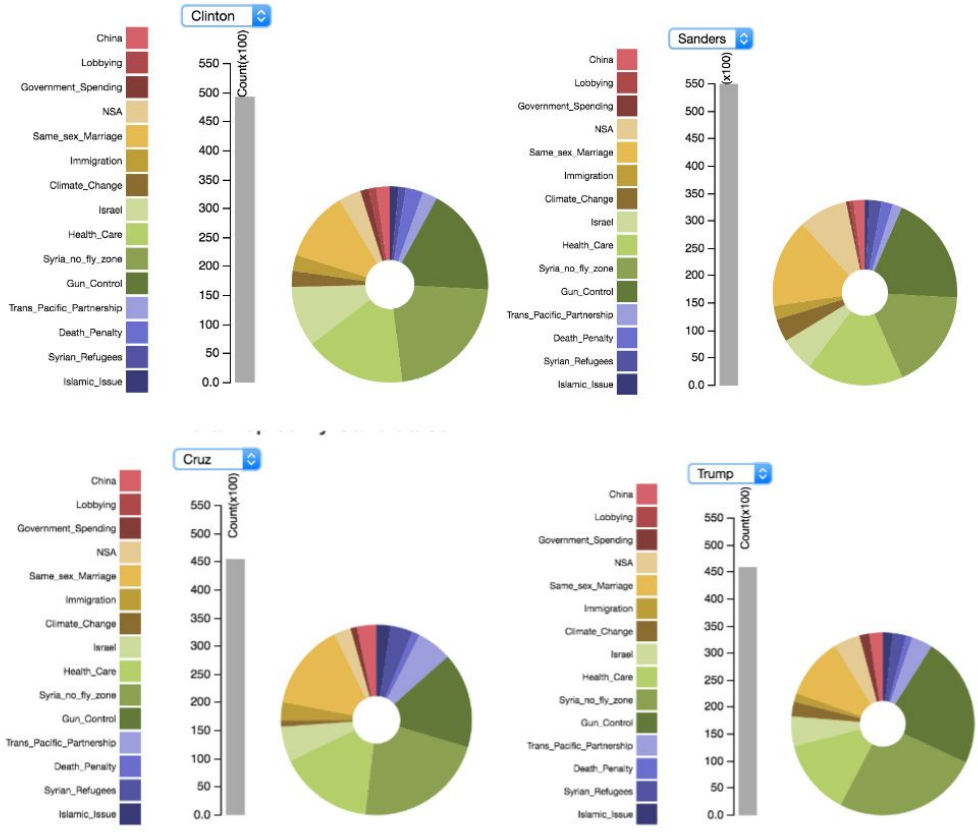


This is interesting to note given what has been said in the media about particular primary candidates focusing a large portion of their campaign on some of these issues (Immigration and 'Islamic Issue''). This discrepancy might be a result of our lexicon but it also might be the result of natural language processing in general and the ability of algorithms to understand subtlety, nuance as well as the sheer amount of words that could be encapsulated into a category like 'Immigration'.

Things to consider:

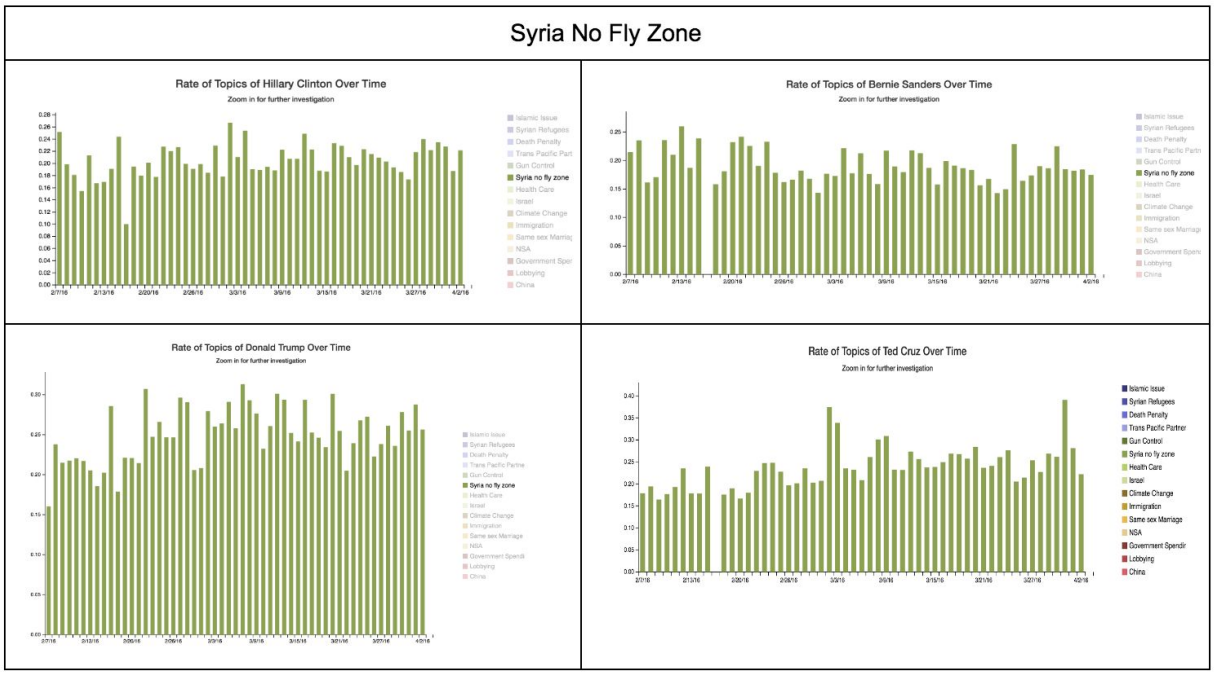
- These topics aren't necessarily the most talked about topics. There could be other topics that our analysis didn't account for (ex. race) that was often sited on Twitter
- Given that this data is coming from tweets, our analysis is not representative of the whole United States. It is interesting and potentially useful to know how each candidate is thought of online.

From the four images below we can see that 'Syria no fly zone', 'Gun Control', and 'Health care' were all major topics that tweets for each candidate included. We can see differences however once we start getting deeper into the data. For example, Cruz's tweets had a larger proportion of 'same sex marriage' tweets while Clinton's tweets had more tweets on 'Israel. Bernie had the largest amount of tweets on the NSA (when compared to the other candidates).



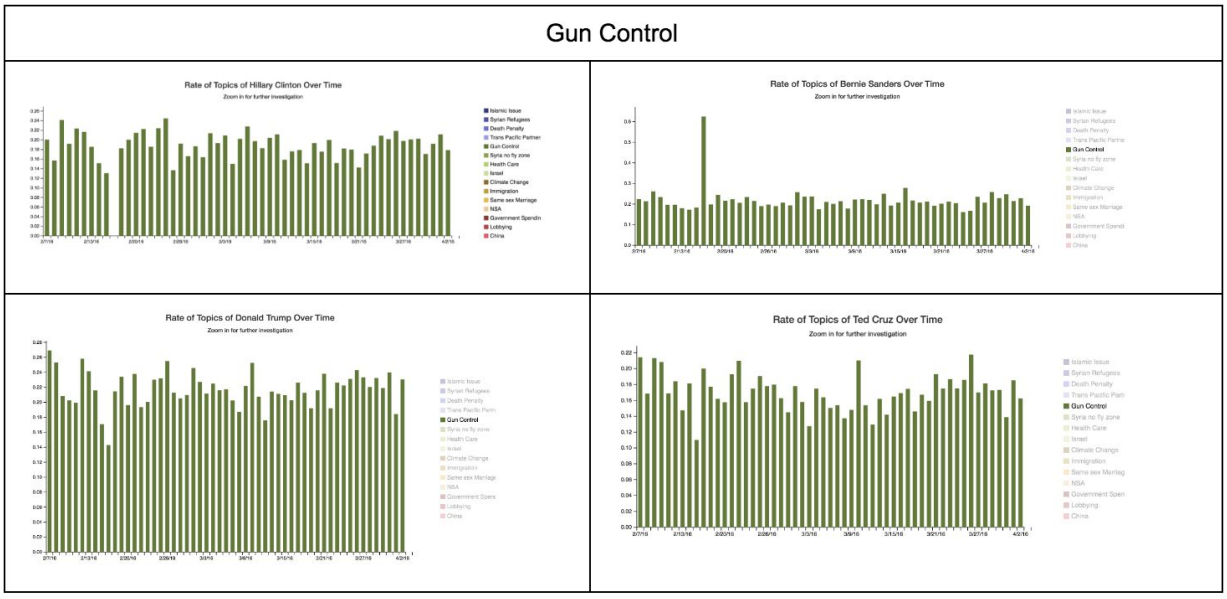
Looking further into issues by day per candidate we can see the following trends:

High percentage overall

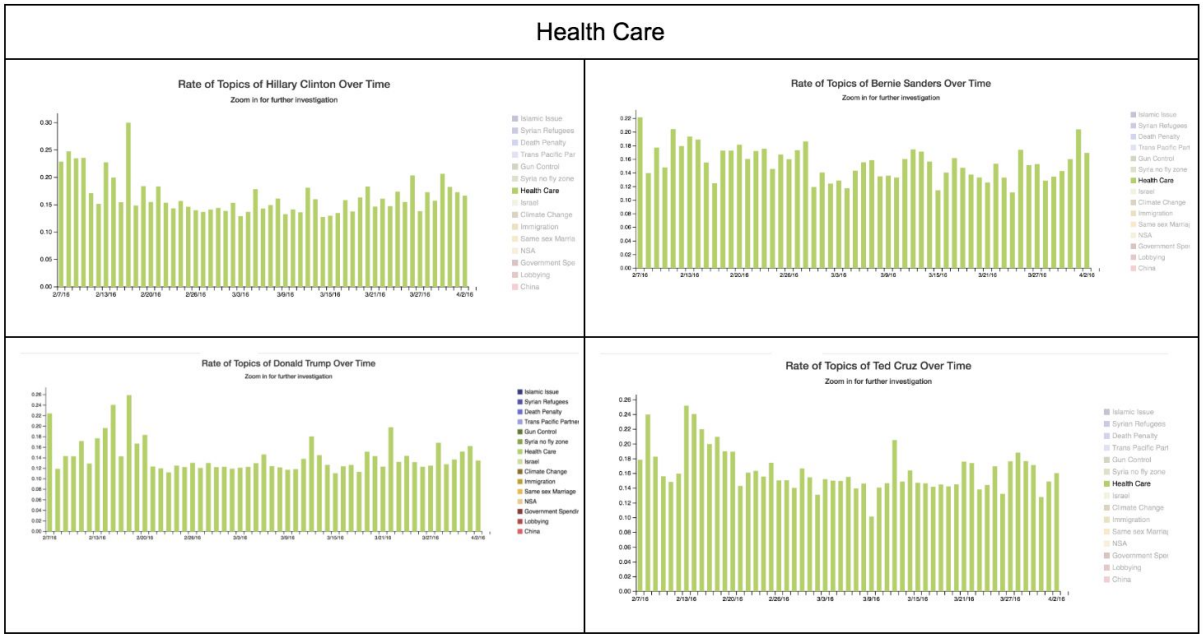


The ‘Syria No Fly Zone’ topic was one of the more heavily mentioned topics. In particular tweets that have been categorized as mentioning Ted Cruz’ have had the highest rates in the category.

A plausible reason could be based on comments given by Ted Cruz in his interview to NBC made a statement, “America has no business sticking his nose in the civil war. In the interview he further substantiated his comments by stating his critical estimation of the US presence or the lack of US presence in Afghanistan, and the dangers associated with the Obama-Clinton foreign policy. Moreover, on being asked about the Syrian issues, his primary stand was to tackle the ISIS, blaming the current administration in its inability to tackle the issue and create a void of power for Russians to explore.



'Gun Control' was also heavily mentioned. Candidates on average mentioned topics related to gun control 20% of all tweets throughout the time range collected. The Second Amendment to the constitution, is a hotly debated issue. The right to bear arms, not only generates is a heated topic for debate and discussion, but also represents polarised opinion across the US, with many states such as Texas, Alabama and Mississippi maintaining a strong pro-gun control stand.

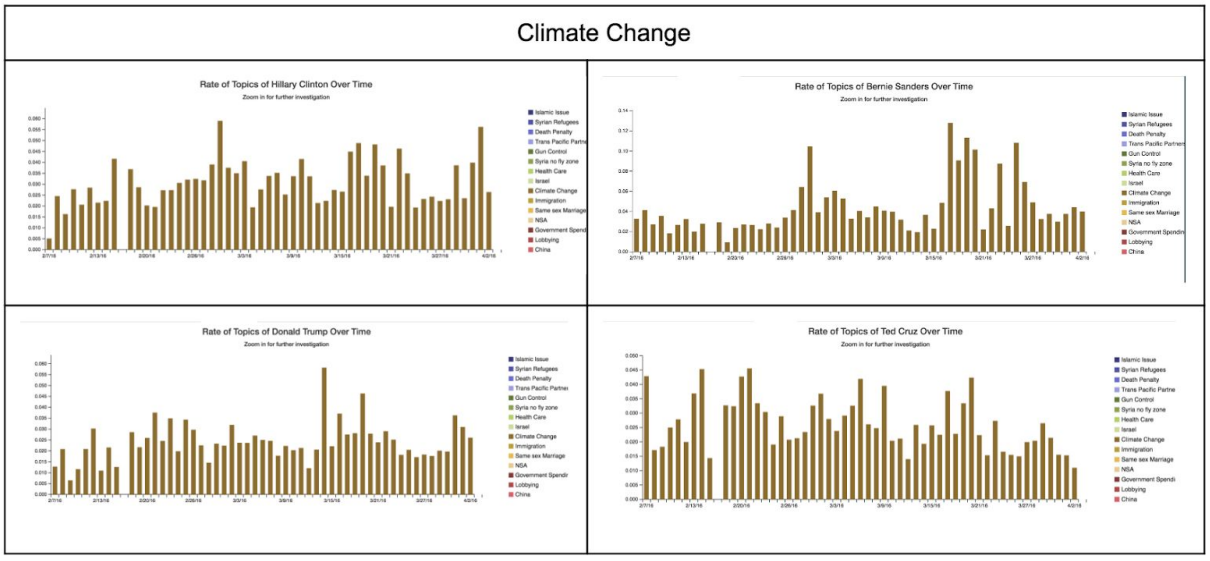


'Health Care' was another heavily mentioned topic. Above we can see, that although it appears that Clinton has a higher mention of topics in this category, the spike in her tweets is on February 17th which was a day that we did not have many if any tweets for that day. As a result, all of the candidates had a similar rate of health-care tweets (max around 25% of daily mentions).

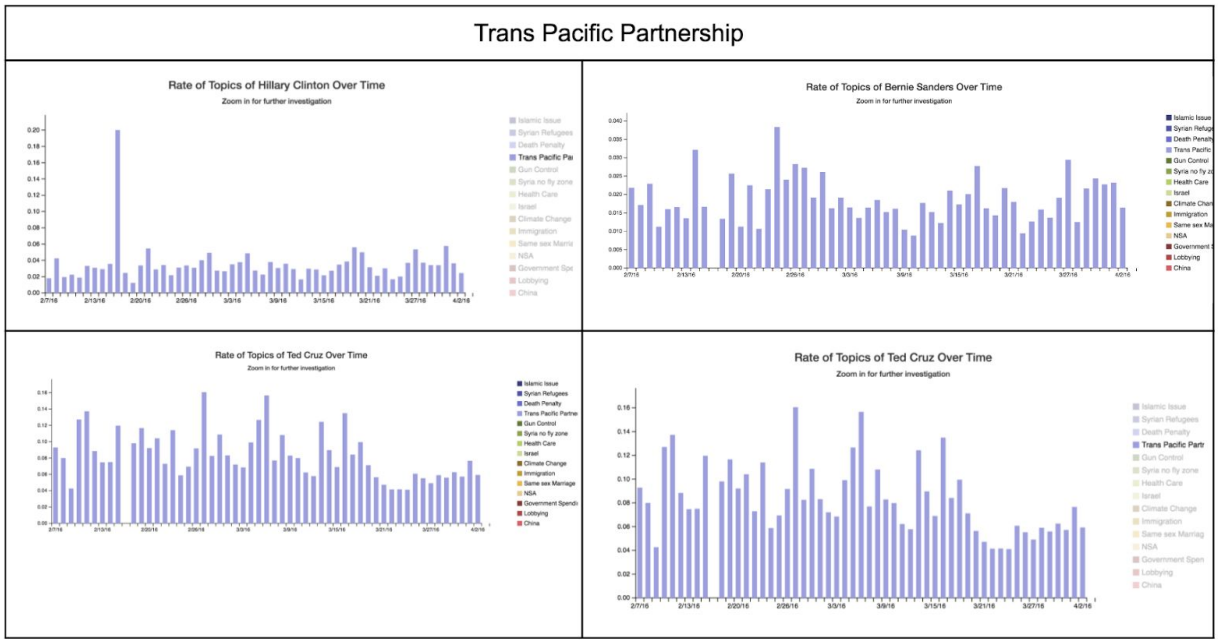
Lastly, 'same sex marriage' was another heavily mentioned topic. Above we can see, though, that Clinton spoke more on average about same sex marriage issues than the other candidates (max 30% of daily tweets). In addition we can see above that compared to the other candidates, Ted Cruz's tweets had the lowest rates for this category (max 16% of daily tweets). Given that Ted Cruz's stance on gay marriage is that "marriage is a sacrament between one man and one woman. It is the building block of civilization and has strengthened societies for millennia. As a constitutional matter, marriage has traditionally been left to the states, and we should protect the rights of elected state legislatures to define marriage consistent with the values of their citizens." [10] We should not be particularly surprised by this.

[10] Taken from Google's Candidates webpage, https://www.google.com/webhp?sourceid=chrome-instant&ion=1&espv=2&ie=UTF-8#q=ted%20cruz%20same%20sex%20marriage%20stance&eob=m.07j6ty//short

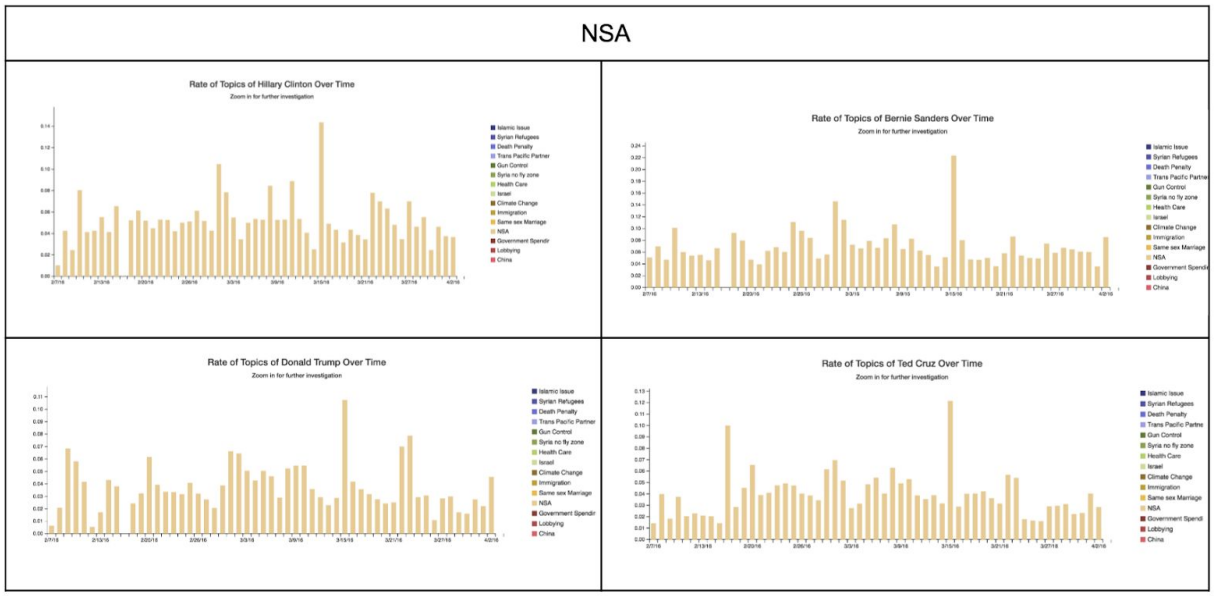
Oddities:



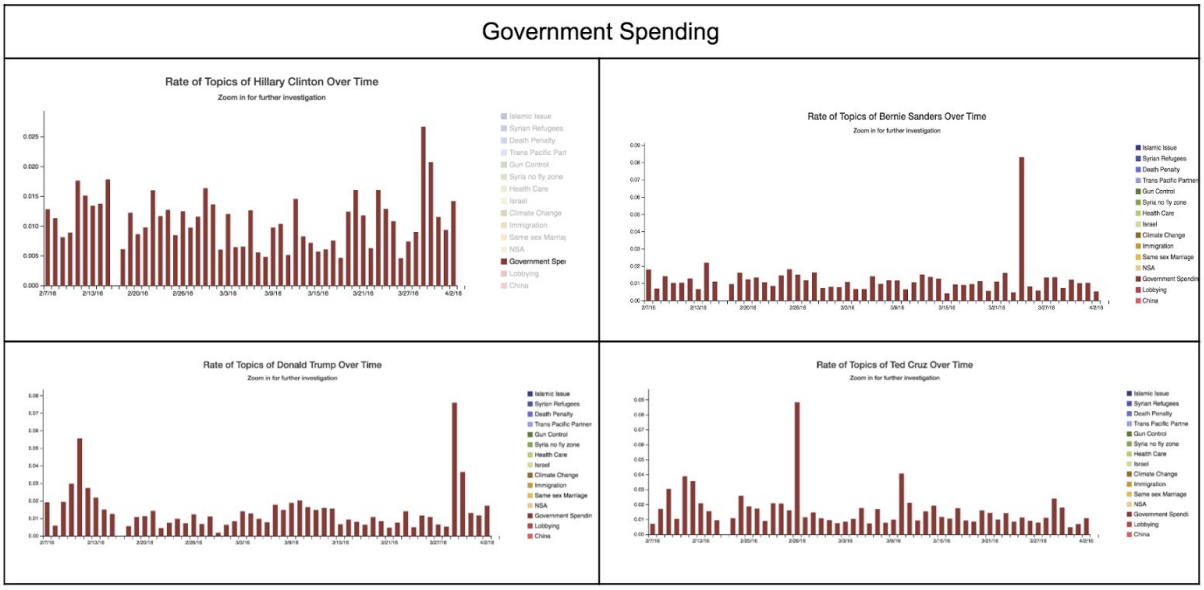
On February 29th Hillary Clinton spoke about Climate Change during the Campaign Rally in at George Mason University in Fairfax, Virginia. On April 1st Hillary Clinton Campaigned in Syracuse, New York In both cases she talked at lengths about the Climate Changed which spiked the interest online and increased the volume of tweets on that topic. On March 14th Donald Trump held a townhall in Tampa, Florida where he expressed his views about the Climate Change thus causing the rise of popularity on that topic.



On February 17th Hillary Clinton campaigned in Chicago where she expressed her views about the Trans Pacific Partnership thus generating a lot of of conversation on the topic.



On March 15th primaries and caucuses took place in 5 states: Florida, Illinois, Missouri, North Carolina and Ohio. The topic of NSA was widely discussed during the previous debates so the interest remained high during the primaries.

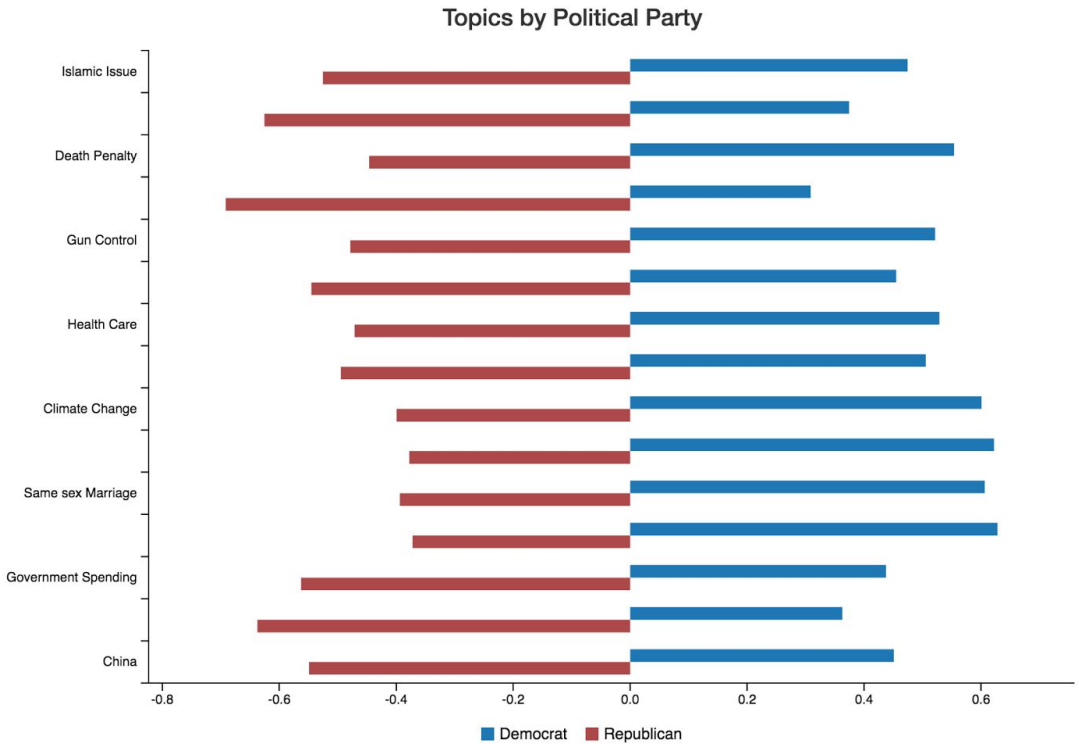


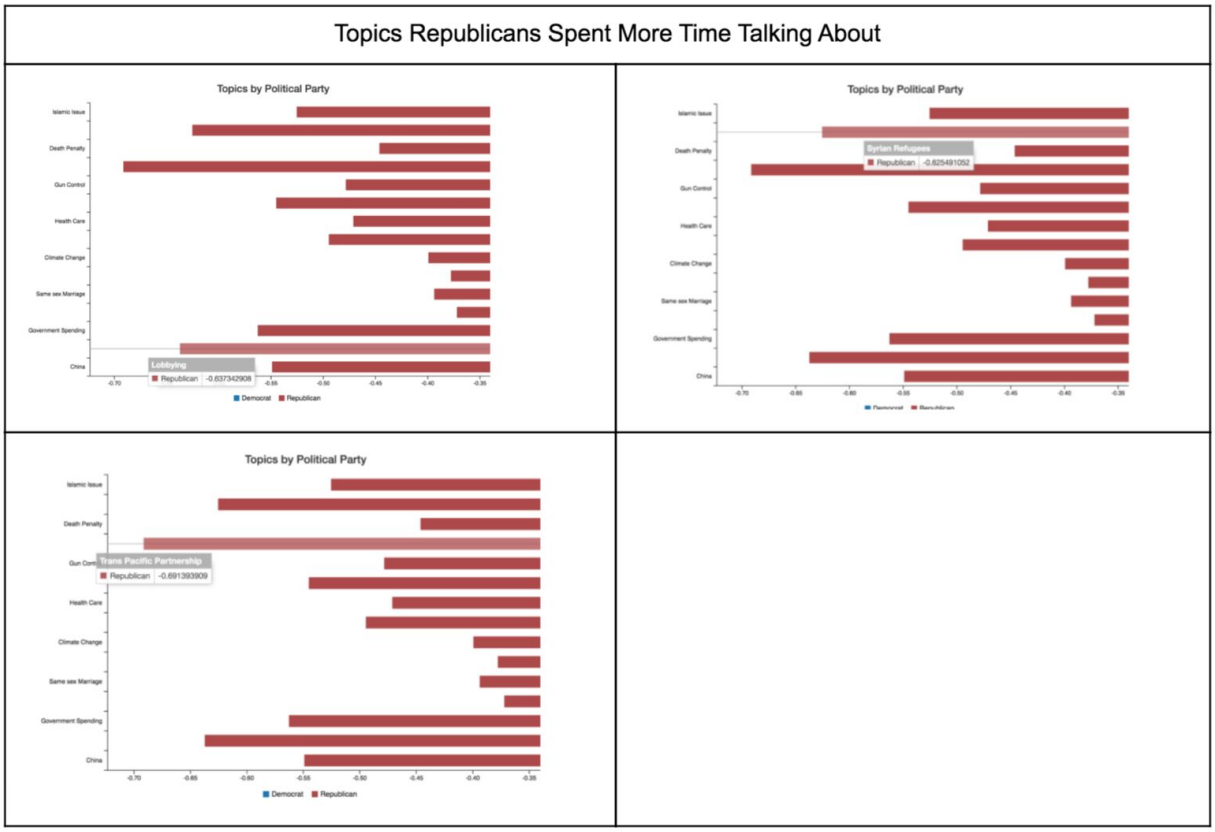
On March 24th Bernie Sanders was talking at a rally in Yakima, Washington. One of the major topics at that rally was the topic of government spending. Which ignited a lot of discussion about the topic on Twitter. On February 26th Republican Nominee Ted Cruz gave an extensive interview to a famous conservative pundit of FOX News - Sean Hannity.[11] In his interview he explained that Obama's economic policies and the government spending are hurting average Americans. This topic was picked up by the Twitter audience and discussed a lot in connection with Ted Cruz.

[11] http://insider.foxnews.com/2016/02/26/ted-cruz-hannity-2016-race-donald-trump-super-tuesday

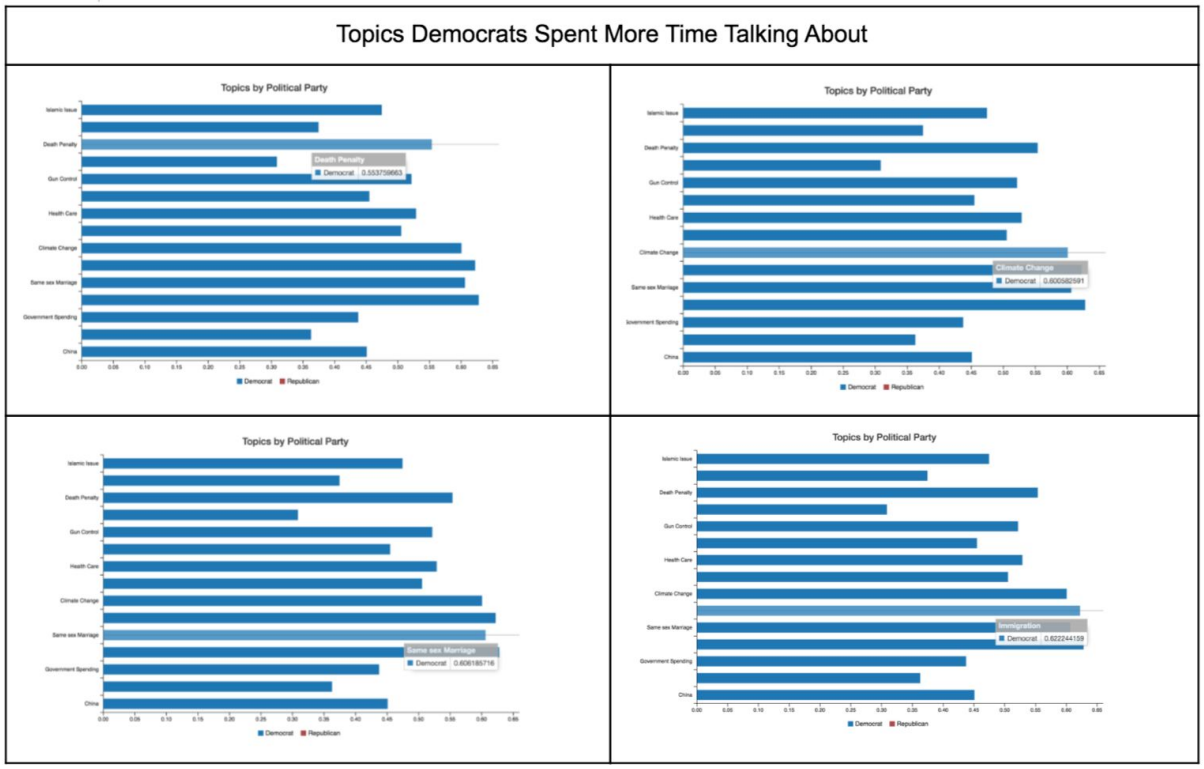
By Party:

Below we can see the breakdown of the topics by political party. The blue bars represent Democrats and the red bars represent Republicans. The topics are calculating the rate of topics spoken by each political party. In the direct image below not all of the topics on the y-axis are displayed however in the four tables below we can see all the topics for each political party.

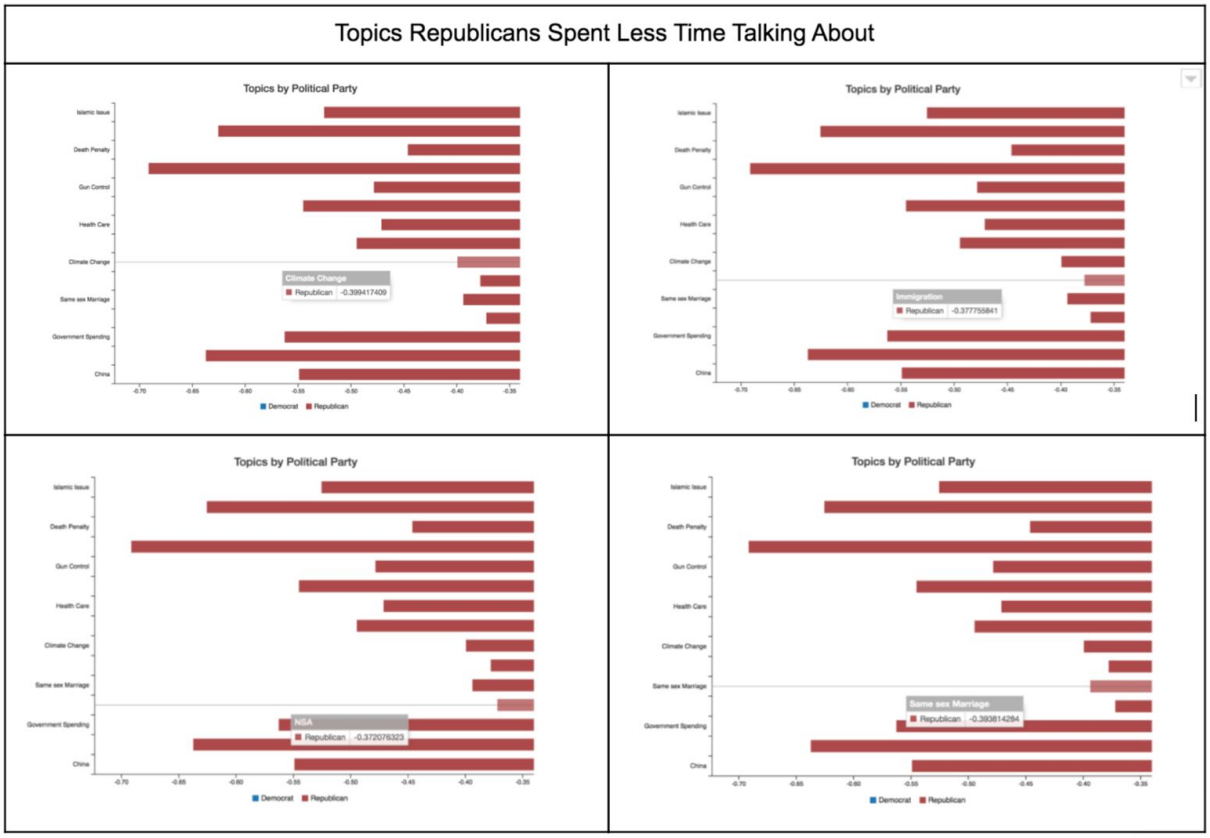




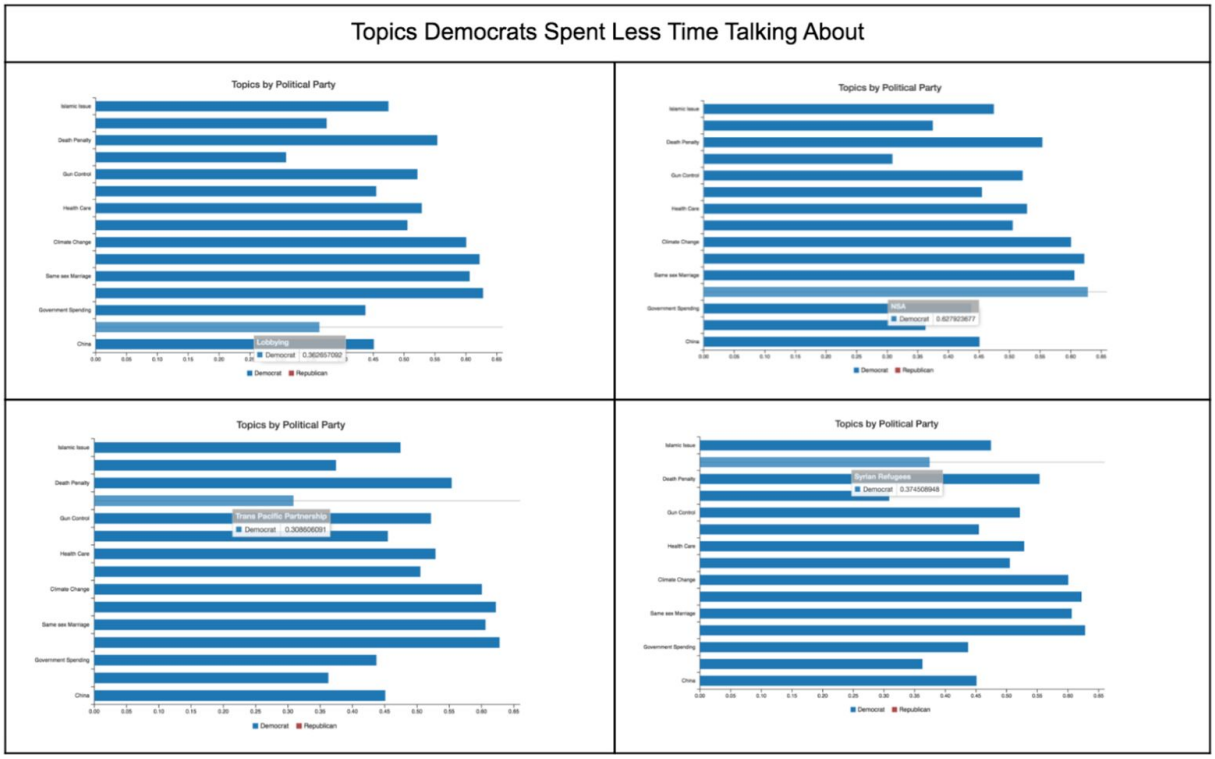
Here we can see that Republicans are 64% more likely to tweet about lobbying, 63% more likely to tweet about syrian refugees and 69% more likely to talk about the trans pacific partnership.



Here we can see that Democrats are 55% more likely to tweet about the death penalty, 60% more likely to tweet about climate change, 61% more likely to tweet about same sex marriage and 62% more likely to talk about the immigration.

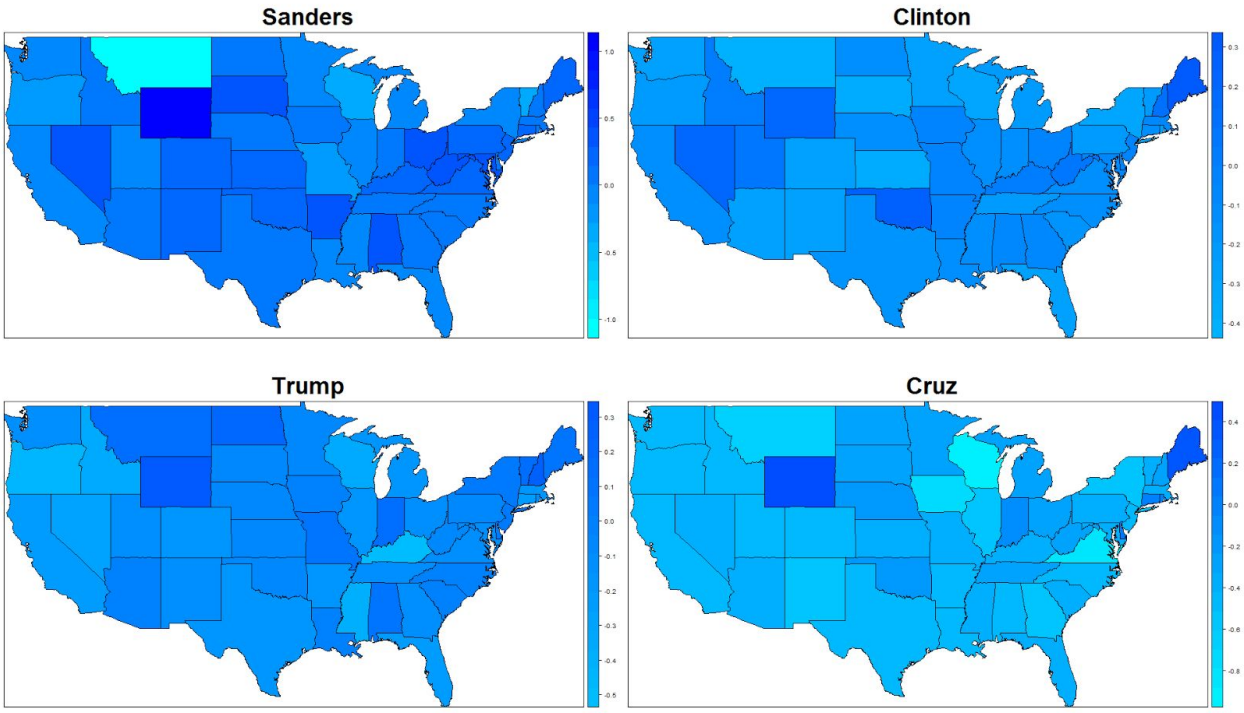


Here we can see that Republicans are only 40% more likely to tweet about climate change, 38% more likely to tweet about immigration, 37% more likely to tweet about the NSA and 39% more likely to talk about same sex marriage.



Here we can see that Democrats are 36% more likely to tweet about the lobbying, 31% more likely to tweet about the trans pacific partnership, and 37% more likely to tweet about Syrian refugees.

MAPPING SENTIMENT OF CANDIDATE TWEETS



The above maps show the sentiment (degree of positivity) of tweets disaggregated by candidates. Before interpreting the maps, we shall briefly describe the methodology of computing the sentiment or degree of positivity. Later we go on to interpret the results and discuss limitations to our methodology.

Methodology:

We initially created a lexicon of positive and negative words by combining the [12]. We only kept unique words in the final version. This was followed by counting the incidence of positive and negative words in the tweets for each state. Initially, we had also included negative and positive emoticons in our analysis but scrapped the idea at a later stage because many emoticons contained metacharacters of regular expressions. We finally computed the following two metrics to measure the degree of positivity:

1. **Positivity index** = Positive / (Negative + Positive). However, the index appeared to be skewed with mean below 0.5 because of much higher negative word counts.
2. For easing interpretation, our team decided to standardize the positivity index between the scale of -1 and 1 (around mean 0). This index is referred to a 'Sentiment ratio score' in the following text and visualizations.

Interpreting the maps:

The intensity of blue in the above maps represents the relative level of positive sentiment across states for various candidates. Since the 'spmap' function in R we used to create the maps, creates a new color scale

[12] https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html

for each map, the color scale of the maps was also standardized so that all candidate maps share the same scale.

The maps show that Sanders and Trump have greater positive sentiment tweets as a whole (greater intensity of blue) as compared to Cruz and Clinton. Cruz tweets appear to the be most negative. Even though Sanders' map appears the most positive at large, surprisingly, he scored the lowest possible scaled score in Montana. On the other hand, Trump scores indicate much more positivity relative to Sanders in the state. Since the volume of tweets was not so large in Montana, to investigate this observation further, our team decided to look at the raw tweets from the state.

We observed that for Sanders, there were tweets that appeared overtly negative such as:

1. "100% true of the #RegressiveLeft @HillaryClinton @BernieSanders"

Moreover, we also came across tweets such as the following example which is not necessarily negative but contributes to negative word counts:

1. "My 23 yr old daughter ' My fucking friends are retards for getting excited that a bird landed on Bernie Sanders podium WTFs wrong w them?"' (An example of sarcasm)
2. "Same nasty tricks against President Barack Obama 2008 are used again toward Bernie Sanders" (seems like a tweet in favor of Sanders but the usage of the word nasty must have contributed to negative word count)

Similarly while observing tweets for Donald Trump, we found the following interesting example which both could be categorized as sarcastic and must have been classified as positive tweets given our lexicon:

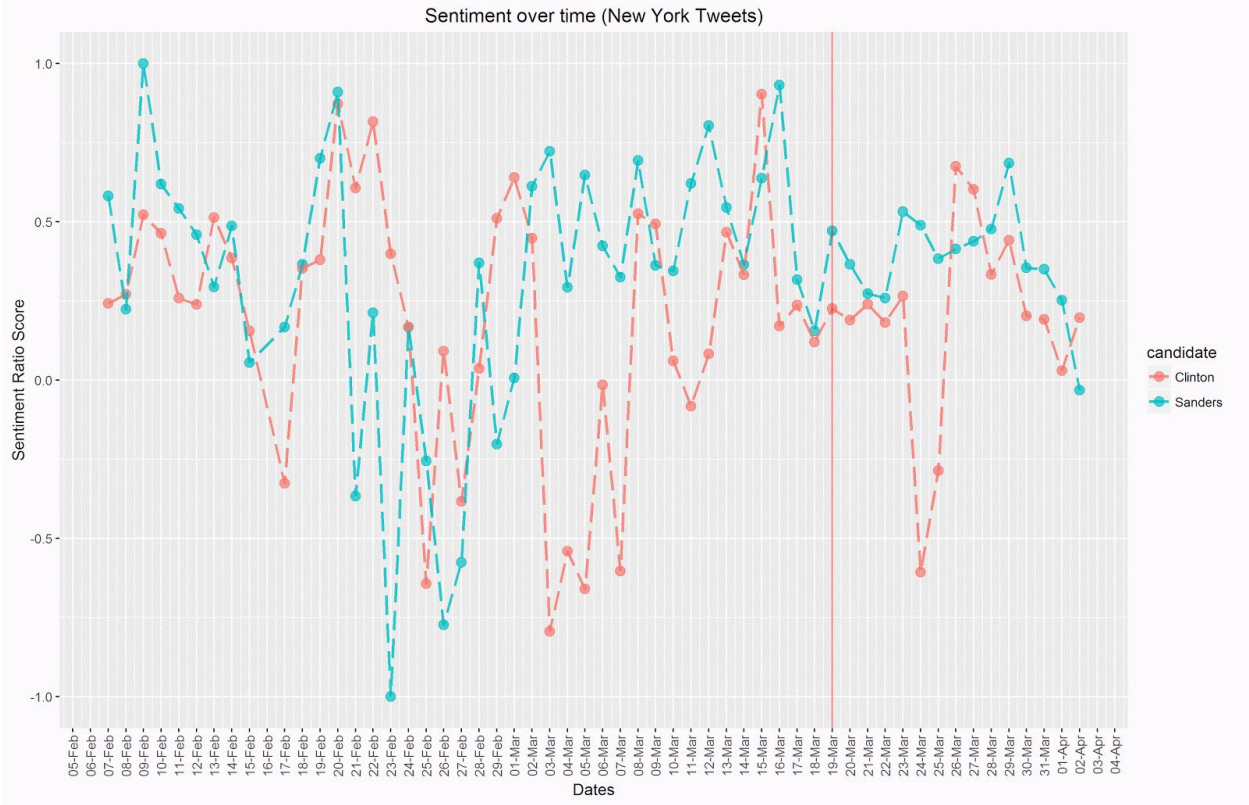
1. "My father said everything Donald touches turns to gold"
2. "The main reason women should vote for Donald Trump . A good chance your husband will get back to work"

The tweets above reflect the limitations of the lexicon based sentiment analysis. The word counts unfortunately miss the very nuances of emotions within sentence structures, making objective sentiment analysis increasingly challenging.

Additionally, given the maps, we also tried to test our assumption of ' whether there is a correlation between state sentiment about a candidate on twitter and the actual primary outcomes, In doing so, we found that surprisingly sentiment ration score calculated by our methodology are not very predictive of which candidates win the presidential primaries. We identified a quite a few states in which our sentiment results are inconsistent with the actual primary outcomes. For example, Sanders' tweets seem more positive in Arkansas and Alabama, despite loss in primaries to Clinton. Moreover, such a discrepancy can be observed for many other states such as in Texas for Cruz vs. Trump in the Republican race.

Time variation in sentiment:

To further evaluate the effectiveness of sentiment as a predictor of primary wins, we decided to study time trends in sentiment ratio scores within a particular state. To this end, we created the following visualization for the state of "New York".



The vertical line highlights the primary date for New York and we can see that Sanders has a slight lead over Clinton (which is contrary to the primary results). In the time window of a few days around the the threshold, we also do not observe any leading patterns. The sentiment within in five days leading to the primary seems to the comparable for both the candidates. This further indicate that even after including the time variable, sentiment ratio scores computed using the lexicon methodology do not seem to be very effective in predicting primary results.

In terms of other interesting observations, we do see a sharp slump for Sanders' sentiment on 23rd of February. We tried to find possible explanations for this anomaly. One of the possible reasons is the political climate after the fiery Wisconsin Democratic debate which was followed by former journalist David Brock accusing Sanders of "negative campaigning"[13]. However, this explanation still remains as a hypothesis that needs to be tested.

Tweet/ Volume as a predictor of primary outcome:

[13] http://www.newsmax.com/Headline/david-brock-bernie-sanders-stop-negative/2016/02/23/id/715785/

Comparing our results for sentiment maps with tweet/ voter population maps that previously developed during class, we observed an advantage of tweets/ volume as a predictor of presidential primary wins over sentiment ratio scores across states (See appendix for standardized volume maps)**.**

CONCLUSION

Given that our main research question was correlating issues of real world events of the US 2016 primaries with the tweets we collected, we see that online social media is able to capture major events that occur in the real world. While the tweet patterns may not be proportional to the patterns generated by real-world events, they allow researchers to be notified of a potential trend and then further investigate - essentially utilizing an exploratory analysis approach. This group also believes that given enough time and data, it is possible to understand real world events or trends happening in real time by looking at the trends and patterns on online social media - in particular Twitter. In order to look into this further it would be useful to have access to other sources of social media like Facebook statuses and microblogs of vlogs. Hence, this paper is an introduction to the social media impact and the election outcomes and concludes on exploring additional tools to draw conclusions whether real world events mimic twitter discussion.

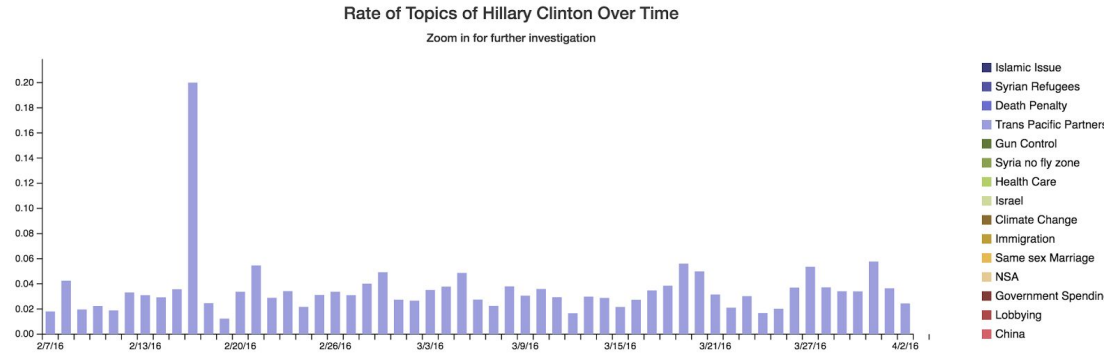
APPENDIX

http://www.usatoday.com/news/
http://www.huffingtonpost.com/entry/marco-rubio-russia-no-fly-zone_us_5612e38ae4b0368a1a60ad06

To access our visualizations visit the following link:
http://www.columbia.edu/~amp2261/D3/index.html
The website displays our D3 code and has a link to our Shiny App as well as our code which is being hosted on github. (see top right hand corners for other links on webpage)

http://www.usnews.com/topics/subjects/gun-control-and-gun-rights

Things to keep in mind:
Because we did rates, some days may appear to have high percentages of a certain topic but if you look at the count of tweets they are quite low. (ex. Feb 17th)



The image below shows topics by political party measured in percentage.

Geo - distribution of Bernie Sanders tweets

Geo - distribution of Hillary Clinton tweets

Geo - distribution of Donald Trump tweets

Geo - distribution of Ted Cruz tweets