Hillary Clinton vs Donald Trump

Sentimental Analysis on Twitter data

Authors

Apoorva Garlanka, Arun Ram Sankaranarayanan , Dwipam Katariya and Jinsu kim

Motivation

When we took the topic about presidential election, it was the United States presidential election season, the question that had been headlining everywhere “who is going to be the next president” all over the world, and people still have interest about how the result came out and different from their expectation. The election of the United States is significant not only to Americans but also to other people all over the world, because the president of the United States has substantial impact on other countries too. The people and the parties are interested to know what public think and their opinions by states-wise. Awareness of the current issues is important to make effective decisions.

Previously, public survey was conducted to know people’s political opinions. However, with the use of the Internet, public opinions can be known from social media. SNS (Social networking sites) is an online service where people build social networks with others who share similar interests, activities, backgrounds or real-life connections. (Social Networking Service, n.d.) With the rapidly growing popularity of SNS, people express their opinions get information from SNS. As presidential election is big issues, users share their political opinions on presidential candidates. In this sense, it has become popular to use SNS to predict future especially presidential election these days.

Twitter is one of the most popular social media platforms in today’s modern world. An aspect of twitter is that it consists rich information about popular trends and happenings around the world. Twitter is a microblogging website wherein tweets are frequently used to express on a subject matter. Twitter has 250+ million active users and generates 500 million tweets everyday. Each tweet can have a maximum of 140 characters in length. People post real time messages about their opinion on variety of issues say reviews, current issues, remarks and so on. The tweets users generate express positive, negative and neutral sentiments. (Twitter, n.d.)

Therefore, we would visualize people’s opinion towards each candidate from twitter data.

Relevant work

With the increasing use of the social media in politics, several studies examined how social media can be sued in election campaign and how this campaign is effective.

Robertson et al., (2010) studied the the use of social media for candidates and voters for the United States presidential election in 2008. He examined the purpose of candidates' use of social media such as Facebook and how this tools can be used for better communication, and how public engage and share information. The researcher analyzed users' pattern of communication of political opinions. In terms of visualization, they mainly focused on statistical analysis using one bubble chart and fourteen bar charts which are simple and easy to recognize, common and traditional analytical visualization. Since it used many bar charts, it is somewhat difficult to see the overall trend and takes time to interpret.

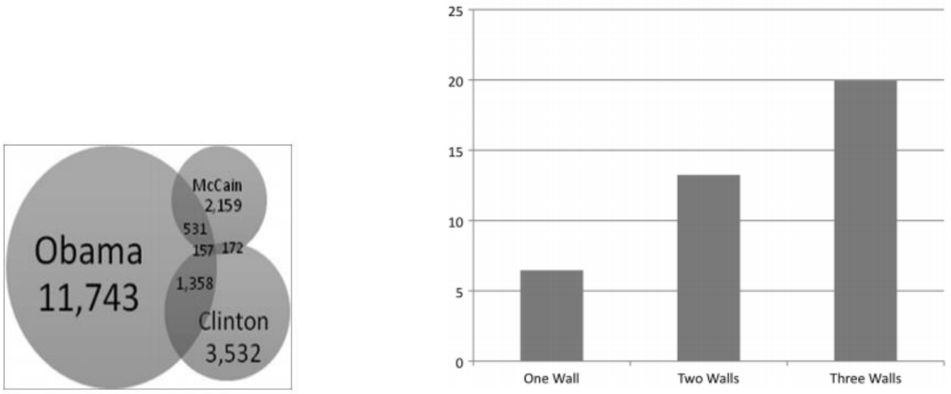


Fig. 3. Number of wall posters who commented on each candidate's wall and across walls (areas approximated).Fig. 4. Average number of posts per poster in the three Wall Crossing conditions.

Fig 1: Bar charts to analyze the trend

Several studies sentiment analysis with twitter. Bermingham and Smeaton (2011) examined how to predict election result and monitored political sentiment using twitter data. They used bar graph to compare prediction and result, and used line chart, sentiment and time series visualization.

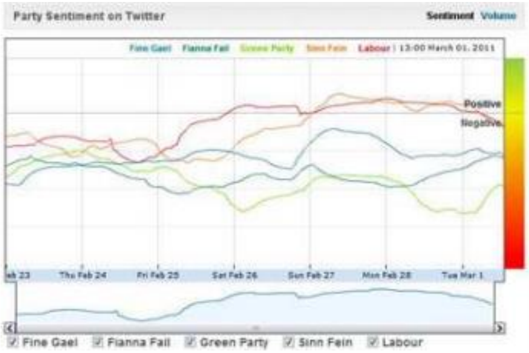
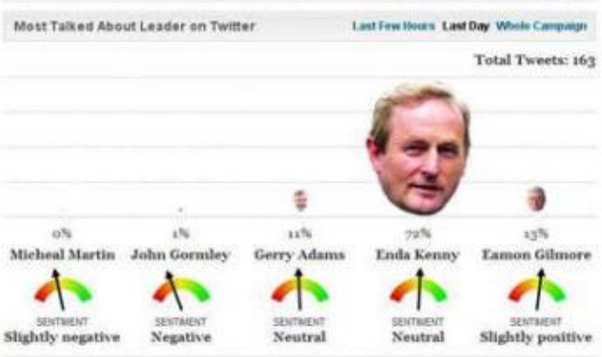
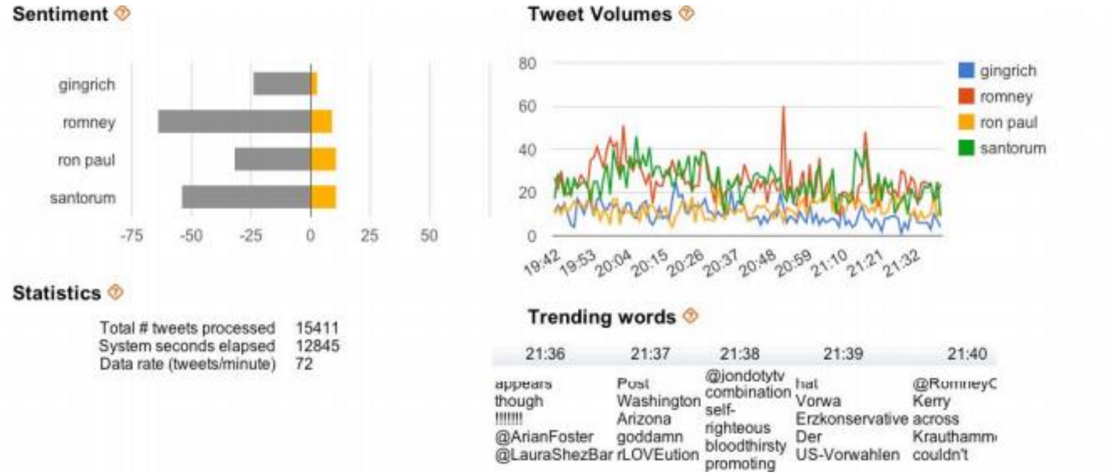


Fig2:



Analyzing the support for political parties and cadidates

Wang et al. (2012) did real-time sentiment analysis using Twitter for the 2012 presidential election. They showed the number of positive and negative tweets about candidates using bar graph which is easy to recognize, and included trending words. To show volume and sentiment about candidates, trending words, they used Ajax-based HTML dashboard that pulls data from a web server every 30 seconds display, which is useful for real-time sentiment analysis. However, the four lines used in the graph are in similar strong color, it would be difficult to see for color blindness, and difficult to recognize in white-black printed paper.

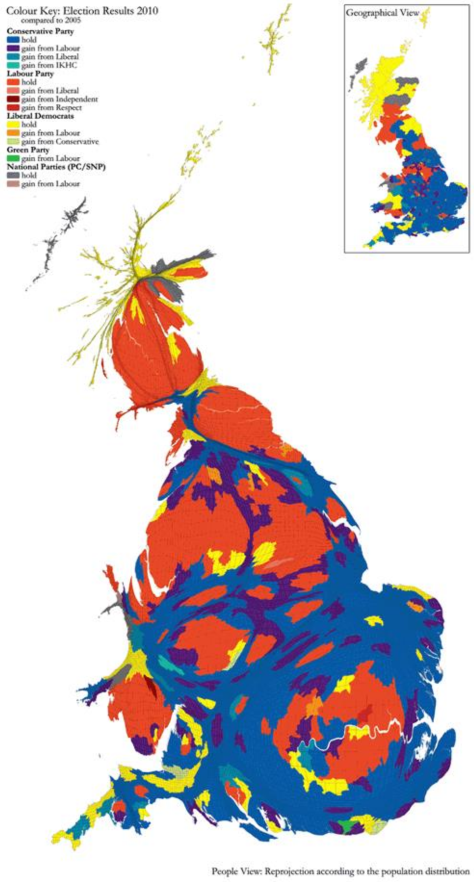


Total # tweets processed	15411
System seconds elapsed	12845
Data rate (tweets/minute)	72

21:36	21:37	21:38	21:39	21:40
appears though !!!!!! @ArianFoster @LauraShezBar	Post Washington Arizona goddamn rLOVEution	@jondotytv combination self- righteous bloodthirsty promoting	hat Vorwa Erzkonservative Der US-Vorwahlen	@RomneyC Kerry across Krauthamm couldn't

Fig 3: Analyzing the support for candidates

There are some studies about presidential election with cartogram visualization. Dorling and Hennig (2010) examined the UK lection result in a cartogram map. The map well showed which parties won and lost. As many other studies did, it visualized in cartogram map with the result data after the presidential election had been done. However, it is also important to figure out which events or debates affect public opinions by visualizing public opinions in real time for better electoral strategies.



Most of the studies about presidential election using social media focused on statistical analysis visualization using simple bar charts or line graph, and the studies took time to process data, get result, and show it. Since opinions about presidential election vary real time due to candidates' debates or campaigns, real-time analysis is more effective. Wang et al. (2012) did the similar study in terms of sentiment analysis of presidential election with real-time twitter data. The study displayed the analysis in real-time, but they did not show users' opinion by location to see each region's opinions.

To infer information easily by region, visualizing public opinions on maps would be effective. Therefore, we will do real-time sentiment analysis for Hillary Clinton and Donald Trump who are candidates of the U.S. presidential election of 2016 and geographic visualization. We would like to know the people's opinion on Hillary Clinton and Donald Trump as whole and state wise. We expect to analyze the positive and negative tweets about the candidates on the worldwide map and visualize by states. With real-time sentiment analysis, it can provide intuition about what events affected public opinions to be negative or positive. With geographic visualization, people can see which states are favor or against to certain candidates. Cartogram map visualization can provide better insights allowing us to figure out how different populations of public have different opinions on a map. By doing so, people can know the overall trend by region and infer the reason of having positive or negative opinions. Ultimately, it can be used for different campaign strategy by different states.

Data & Methods: ideas, sketches, prototypes

Preliminary idea

The preliminary idea was to extract the data from twitter and then do sentimental analysis on the data. Our idea revolved around presidential elections in particular. We wanted to extract data based on the presidential candidate Hillary Clinton and then perform sentimental analysis on the tweets to intuitively estimate whether there is positive or negative response for Hillary Clinton among the general public.

Hypothesis

After contemplating on our initial idea we expanded our scope to extracting tweets for both Donald Trump and Hillary Clinton. Our hypothesis was to achieve two main goals

- Analyzing and comparing the likes and dislikes that people show towards Hillary Clinton and Donald Trump
- Visualizing the likes and dislikes in a geographical map (cartogram) to analyze the opinion and volume of people liking and disliking in each state of the United States.
- Visualizing the most frequent words associated with Hilary Clinton and Donald Trump.

Visualization methods

The data that is being used for our visualization is twitter data. A systematic model was developed to help understand the different tasks that needed to be performed en route to visualizing the data. The process flow model is being described below.

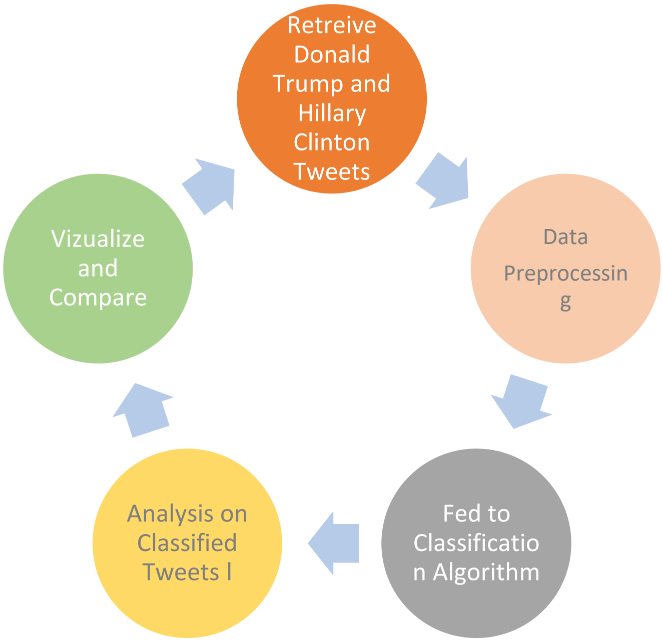

Retreive Donald Trump and Hillary Clinton Tweets
Data Preprocessin g
Fed to Classificatio n Algorithm
Analysis on Classified Tweets l
Vizualize and Compare

Fig 4; Process flow model

The first step that is being done is retrieval of tweets from Twitter. The retrieval of tweets is being done through a call to a twitter API and tweets are being retrieved based on their reference to Hillary Clinton (@HillaryClinton) and Donald Trump (@DonaldTrump).

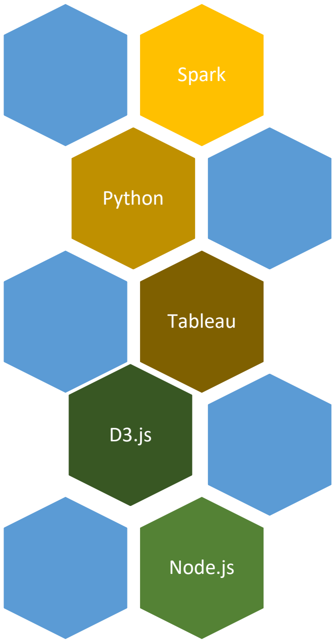
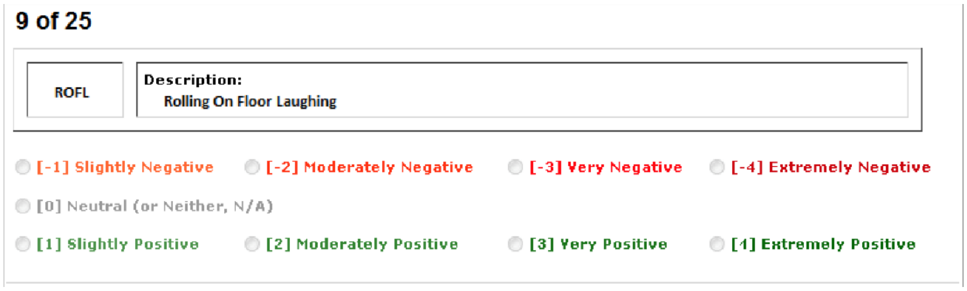


Fig 5: Tools Used

The second step is to remove the stop words and the text that are not needed. The stop words do not actually reflect the sentiments of the people and they can be ignored. There were lots of URLs that were mentioned in the tweets, these URLs can also be ignored as they do not affect our analysis in any way.

The third step is where the tweets that have been retrieved and made free of stop words are being fed into a classification algorithm for the purpose of sentimental analysis. Here NLTK platform is used for performing Natural Language Processing on our data. Using this the data that has been retrieved has been labeled as one of either positive, negative or neutral classes. For sentiment analysis, we use Vader algorithm. Because, Vader outperforms other algorithms such as Naïve Baye's, HMM for sentimental analysis. As, Vader is the parsimonious rule based modeling technique, where it generates rule based lexicons to rate the sentiments in positive, neutral and negative classes. A semantic lexicon is the list of words that are generally labeled as positive or negative. And manually creating train data is the most robust approach. Vader method of NLTK package provides us with already trained model. They collected the data from Amazon Mechanical Turk, where all the users willing to work on the task are given tweets and asked to rate a score for positive, negative and Neutral. After that each tweet was computed for mean and standard deviation of the scores to remove randomness. After that they train different rule based classifiers. We then use these train model to classify the tweets.



9 of 25

ROFL

Description:
Rolling On Floor Laughing

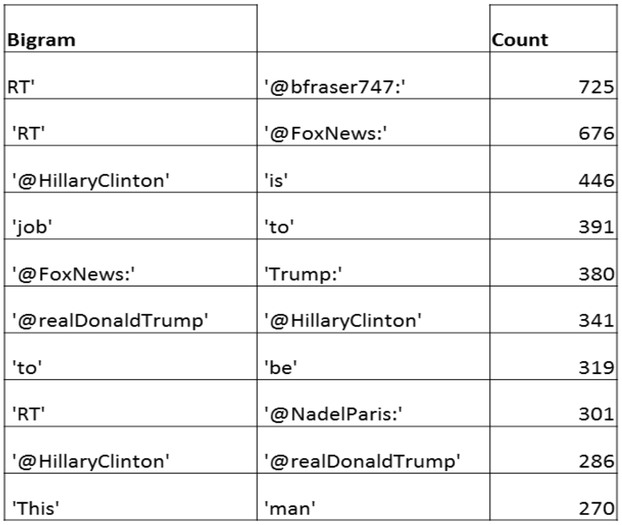
○ [-1] Slightly Negative ○ [-2] Moderately Negative ○ [-3] Very Negative ○ [-4] Extremely Negative

○ [0] Neutral (or Neither, N/A)

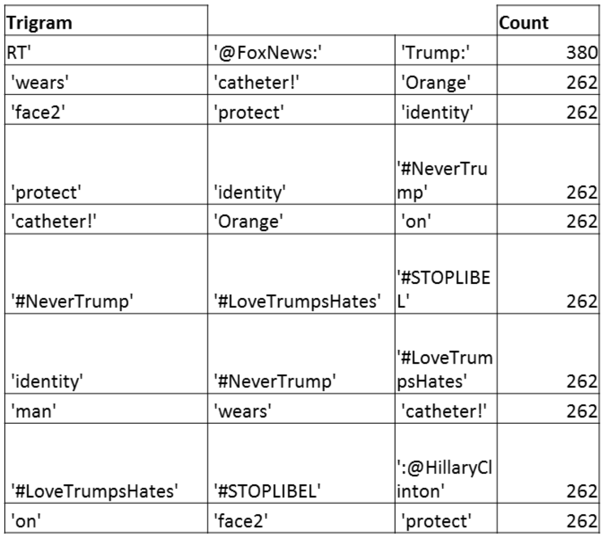
○ [1] Slightly Positive ○ [2] Moderately Positive ○ [3] Very Positive ○ [4] Extremely Positive

Fig 6:Image from Amazon Mechanical Turk to collect the data.

The fourth step would be to analyze on the classified tweets. Based on the classified tweets further analysis have been made to get a clear idea on what the people mostly talk about when talking about Hillary Clinton or Donald Trump. N gram analysis was made on the data collected about Hillary and the results of the bigram and Trigram have been displayed below



Bigram		Count
RT'	'@bfraser747:'	725
'RT'	'@FoxNews:'	676
'@HillaryClinton'	'is'	446
'job'	'to'	391
'@FoxNews:'	'Trump:'	380
'@realDonaldTrump'	'@HillaryClinton'	341
'to'	'be'	319
'RT'	'@NadelParis:'	301
'@HillaryClinton'	'@realDonaldTrump'	286
'This'	'man'	270



Trigram			Count
RT'	'@FoxNews:'	'Trump:'	380
'wears'	'catheter!'	'Orange'	262
'face2'	'protect'	'identity'	262
'protect'	'identity'	'#NeverTrump'	262
'catheter!'	'Orange'	'on'	262
'#NeverTrump'	'#LoveTrumpsHates'	'#STOPLIBEL'	262
'identity'	'#NeverTrump'	'#LoveTrumpsHates'	262
'man'	'wears'	'catheter!'	262
'#LoveTrumpsHates'	'#STOPLIBEL'	':@HillaryClinton'	262
'on'	'face2'	'protect'	262

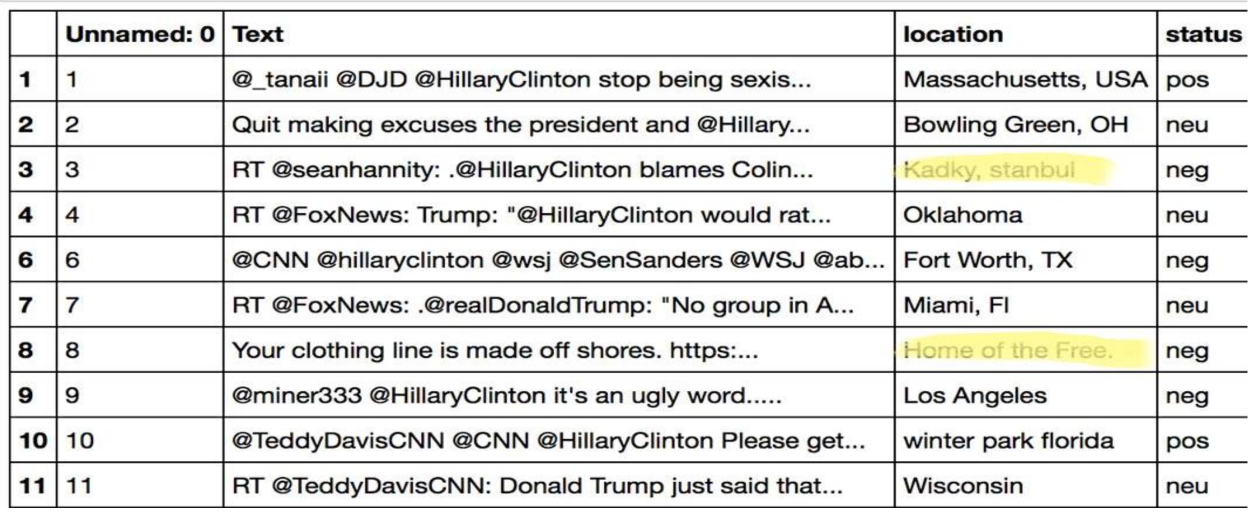
Fig 3: Trigram and Bigram Analysis

Once the Trigram and Bigram Analysis have been identified it is very essential for us to compare the positive, negative and the neutral tweets. For our purpose of visualizing our classified tweets we used bar charts. The main advantage of using the bar chart is that it would give us quick and easy visualization of which word is the most used word under each of the classified tweets category.

We used Geo-Visualization to visualize the classified tweets based on their locations. This visualization was intuitive as it graphically gave us exact location and number of tweets information based on the location.

One other visualization that we used was to depict the volume of likes and dislikes for Hilary Clinton and Donald Trump in an interactive web based application that contained Cartogram. We chose cartogram because each region has different population densities depending on the USA presidential election and we wanted to represent the public opinion with each state’s true weight with cartogram.

Apart from the geographical visualizations, we created a word cloud and force directed graph to see what words occur the most frequent for Hilary Clinton and Donald Trump and the words that are generally associated with each other.



	Unnamed: 0	Text	location	status
1	1	@_tanaii @DJD @HillaryClinton stop being sexis...	Massachusetts, USA	pos
2	2	Quit making excuses the president and @Hillary...	Bowling Green, OH	neu
3	3	RT @seanhannity: .@HillaryClinton blames Colin...	Kadky, stanbul	neg
4	4	RT @FoxNews: Trump: "@HillaryClinton would rat...	Oklahoma	neu
6	6	@CNN @hillaryclinton @wsj @SenSanders @WSJ @ab...	Fort Worth, TX	neg
7	7	RT @FoxNews: .@realDonaldTrump: "No group in A...	Miami, Fl	neu
8	8	Your clothing line is made off shores. https:...	Home of the Free.	neg
9	9	@miner333 @HillaryClinton it's an ugly word.....	Los Angeles	neg
10	10	@TeddyDavisCNN @CNN @HillaryClinton Please get...	winter park florida	pos
11	11	RT @TeddyDavisCNN: Donald Trump just said that...	Wisconsin	neu

fig5: Image of the data captured after NLP

The final data after all the preprocessing had all the states with the volume of positive, negative and neutral counts. This data was used to build the cartogram. We arrived with two tables one for Hilary Clinton and one for Donald Trump

Pros

Some of the reasons why we think our visualizations effectively communicated our data are:

- Bar chart is very readable and gives a quick and easy visualization of which word is the most used word under each of the classified category.
- Geo-Visualization helps us visualize the classified tweets based on their locations very effectively. It is one such visualization that can be understood and draws any viewers attention.
- Cartogram is used to visualize each region in its true weight and not be biased towards different population densities
- Word Cloud is used to deliver the results in a fast and easy way. They are also very engaging.
- Force directed graph may correspond to some meaningful analysis on the source,frequencies.

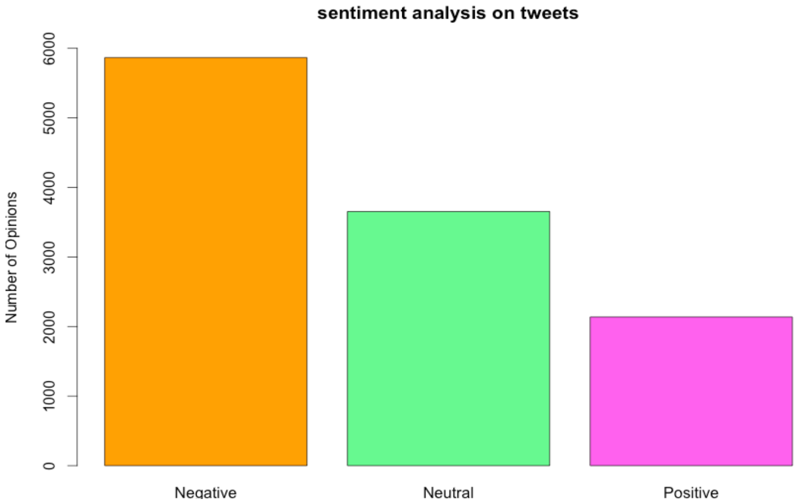
Cons

Some of the cons in our visualization methods have been analyzed below

- Bar charts cannot provide exact quantitative information when there are millions of tweets. It would be very hard to quantitatively differentiate our analysis
- Geo-visualization that we provided will not provide elementary information like name of the county or a certain specific location, we will have to ignore the tweets outside the united states to obtain such clear information.
- Word cloud cannot give a rank for the most frequent occurred word. It's hard to compare which word is repeated more than the other and the information on how many times the word occurred is not clear.
- Cartogram shrinks the the map as per density but fails to give a clear picture when the difference in the volume is not significantly large enough.
- All the visualizations are not assorted in a single platform. Hence the viewer may find it hard to correlate one visualization to the other.

Results

Novelty and Insights

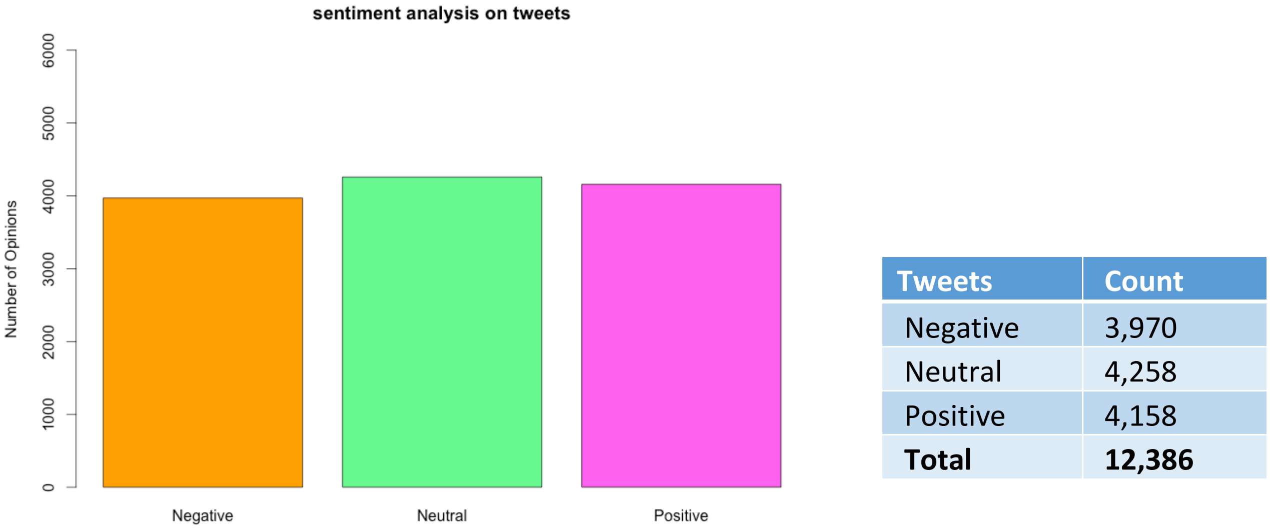


Tweets	Count
Negative	5,866
Neutral	3,633
Positive	2,138
Total	**11,657**

Fig 6: Positive,Negative and Neutral overview (Hillary Clinton) & Dataset Summary

From the tweets that were retrieved based on Hillary Clinton we could see that there was more negativity expressed from our sample. Since this is only a sample we would not be able to concretely say that more people had a negative opinion about Hillary Clinton. This only provides us a bird's eye view on the people's opinion about Hilary Clinton.

The locations from which these tweets were tweeted were quite varied. There was a diverse distribution of tweets in different locations. From the sample that we considered there were close to 4000 locations from which people tweeted tweets about Hillary Clinton.



Tweets	Count
Negative	3,970
Neutral	4,258
Positive	4,158
Total	**12,386**

Fig 7: Positive,Negative and Neutral overview (Donald Trump) & Dataset Summary

From the tweets that were retrieved based on Donald Trump, we can see that almost equal number of neutral and positive opinions, and a little smaller number of negative opinions.

Geo Visualization

Based on the data that we got for Hillary Clinton and Donald Trump we were able to geographically visualize the data geographically based on the amount of positive and the negative tweets. Since the presidential candidate’s role is mostly influenced by the positive and negativity spread about the candidates we narrowed down our results to the amount of positive and negative tweets tweeted about the presidential candidate. Since Cartogram iss one effective geo-visualization technique based on the density of data we stuck to using cartograms on our data.

Data Transformation

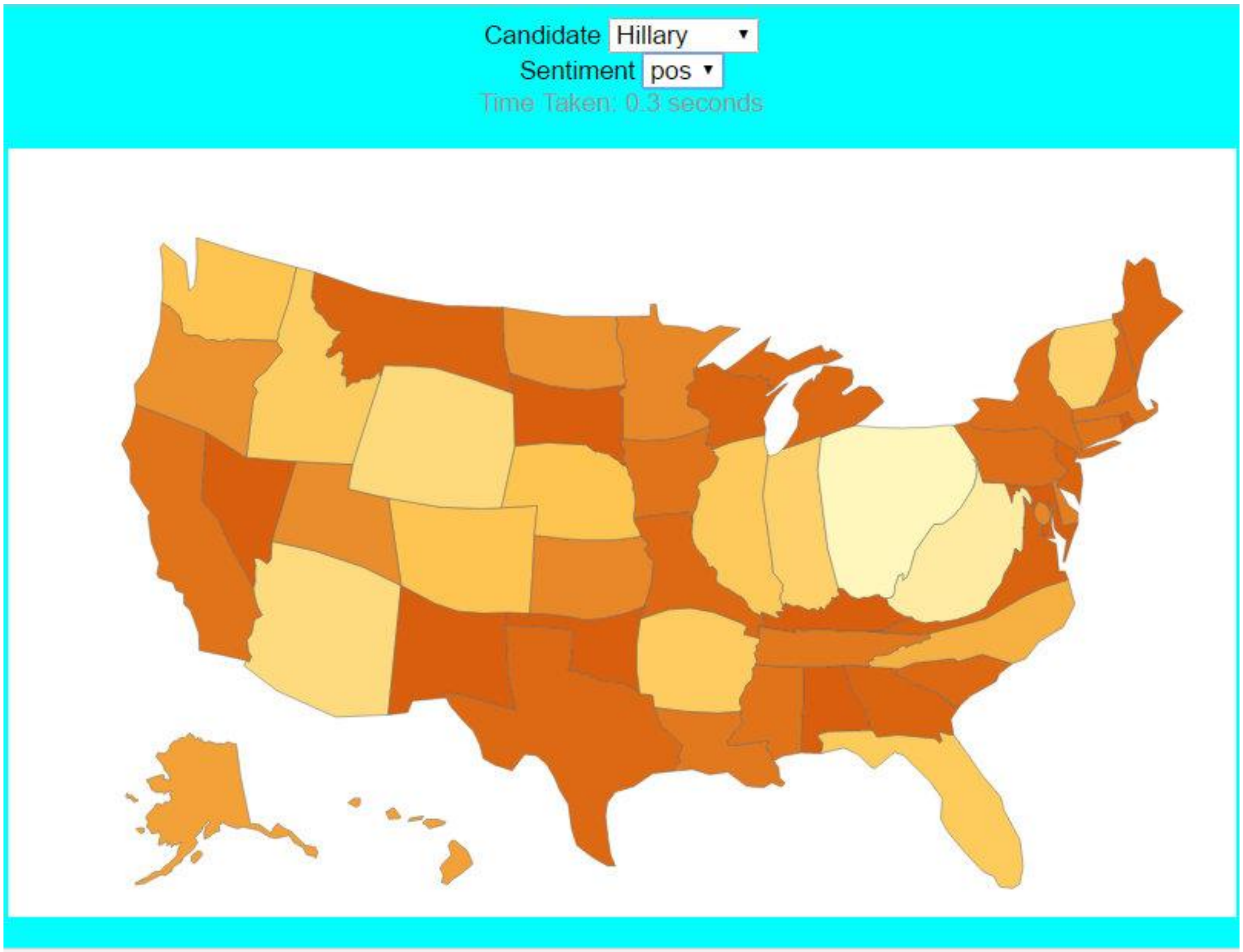
Since the data we collected was based on various geographical locations it was important for us to narrow down to the geographical locations which mainly would affect the results of the U.S presidential elections. The following are the steps we followed to better transform the data

- The tweets that we collected had locations from all over the world tweeting about Hillary Clinton and Donald Trump. We filtered the tweets to only concentrate on locations based on U.S.
- Since our data was based on counts of the amount of positive and negative tweets it would be better for us to use a transformation that would best project the data. We could not work on log transformation since the data did not follow a normal distribution. Hence we stuck to square root transformation and represented the data based on the square root of the data on various states. This helped us to better project our data.

Cartograms

We used cartograms on the transformed data to analyze and understand Hillary Clinton and Donald Trump’s popularity in the United States based on tweets The Cartograms used color patterns to intuitively understand the amount of positivity and negativity in each state. In our data we resort color patterns with different hues to better represent the density of data.

Here to represent lower density in data we stick to darker color shades and to represent high density in data we resort to lighter shades of hue. With this we would be able to intuitively analyze the amount of positive and negative tweets for each of the presidential candidates based on the geographical locations.



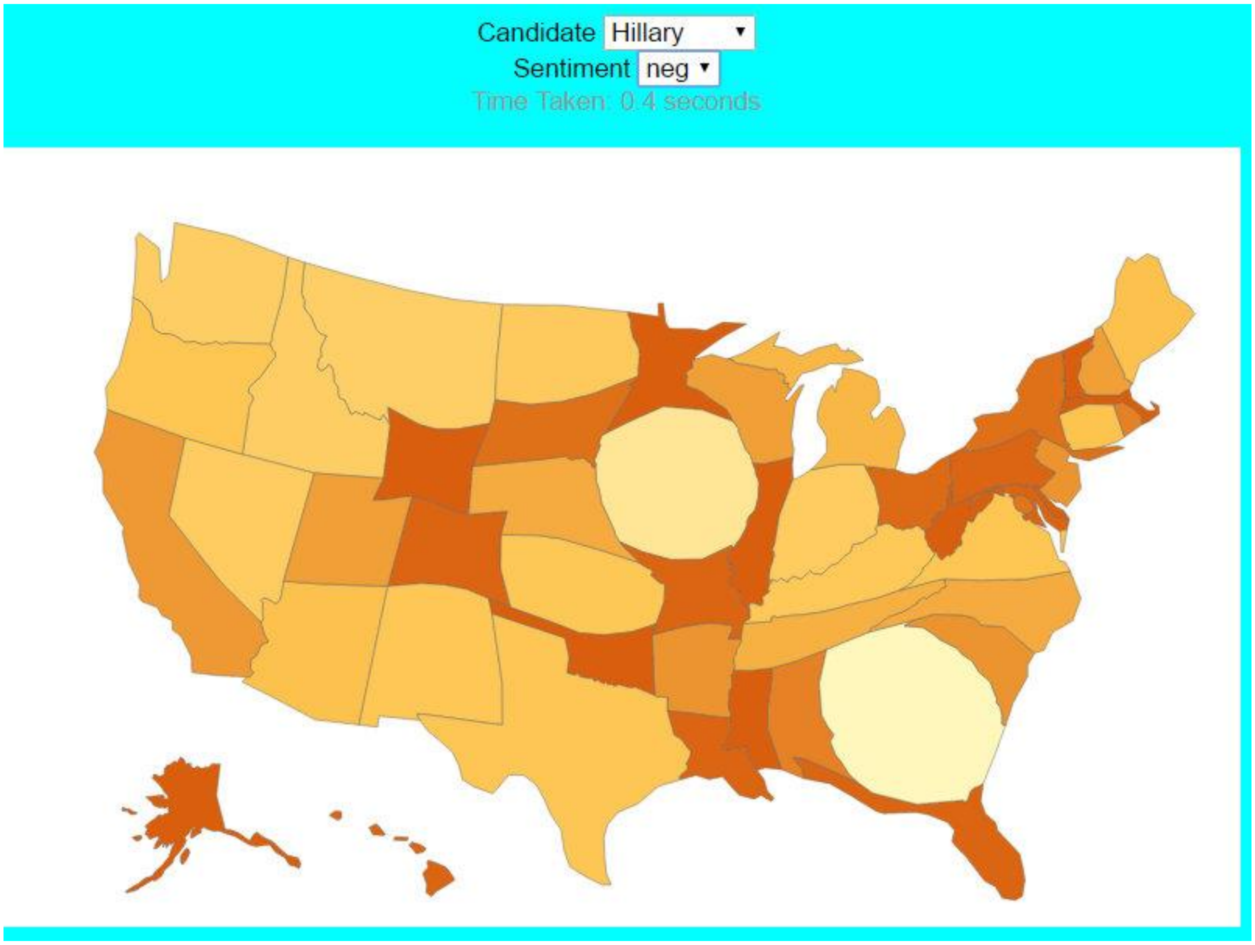
Candidate Hillary ▾

Sentiment pos ▾

Time Taken: 0.3 seconds

Fig 8 : Cartogram representing the amount of positive tweets for Hillary in US

The data effectively showed the amount of positivity exhibited about the presidential candidate Hillary Clinton. The data had comparatively high amounts of positivity in states where she eventually lost. Some of the results here did not reflect the reason for her loss



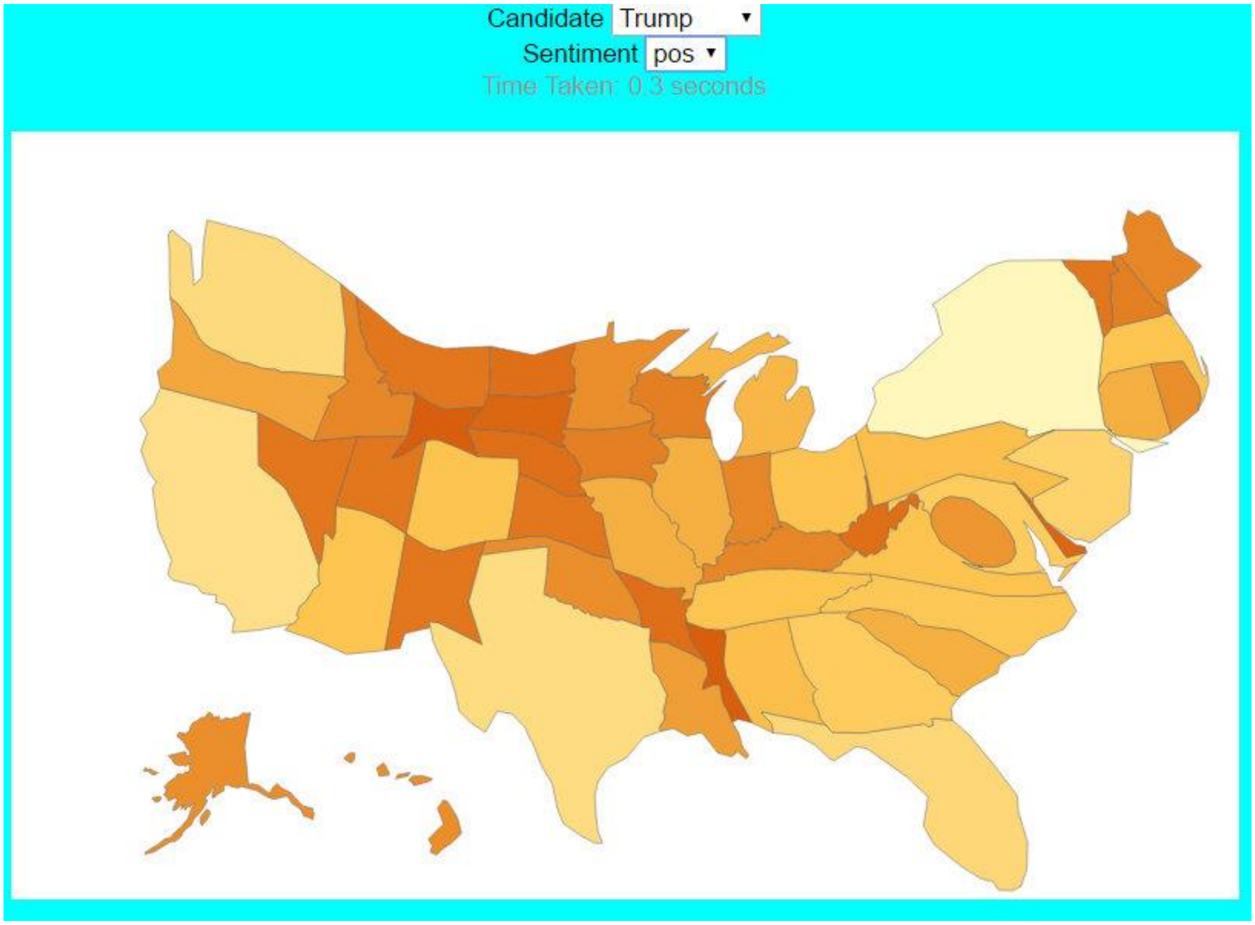
Candidate Hillary ▾

Sentiment neg ▾

Time Taken: 0.4 seconds

Fig 9 : Cartogram representing the amount of negative tweets for Hillary in US

This cartogram effectively showed how there were more negative tweets in which Clinton had less popularity. Some states like Georgia, Iowa and Montana showed considerable increase in the amount of negative tweets. This showed that there were more amount of negativity amongst the people who tweeted about Hillary in these states. These were eventually the states in which Hillary lost.



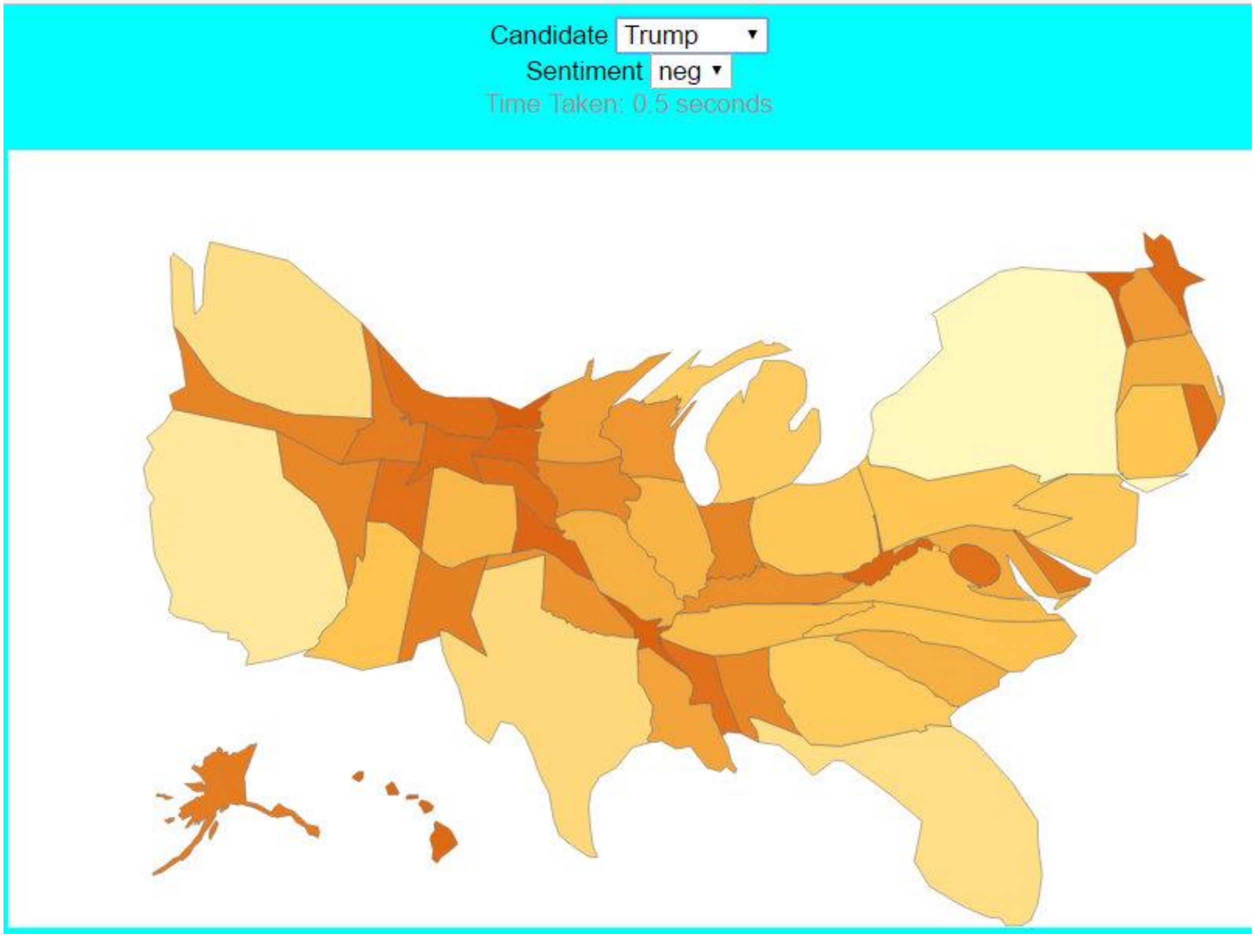
Candidate Trump ▾

Sentiment pos ▾

Time Taken: 0.3 seconds

Fig 10: Cartogram representing the amount of positive and tweets across the state for Trump

The above cartogram shows the amount of positivity Donald Trump has in some of the states in the US. Some of the states like Texas , florida and Georgia exhibited more positivity. This showed that most people tweeted more and supported more for Trump in these states. These were some of the states where Trump eventually won by a huge margin after we came to know about the election results



Candidate Trump ▾

Sentiment neg ▾

Time Taken: 0.5 seconds

Fig 11 : Cartogram representing negativity in tweets

From the cartogram we could clearly see that Trump had huge negativity in the tweets tweeted by people from Washington, New York and California. These are some of the states which eventually turned out to be the states he lost in the elections.

The visualization of the cartogram gave us quite a few insight about the positivity and negativity for each of the candidates in each of these states. This helped us to effectively compare the effectiveness of our visualization.

Word cloud

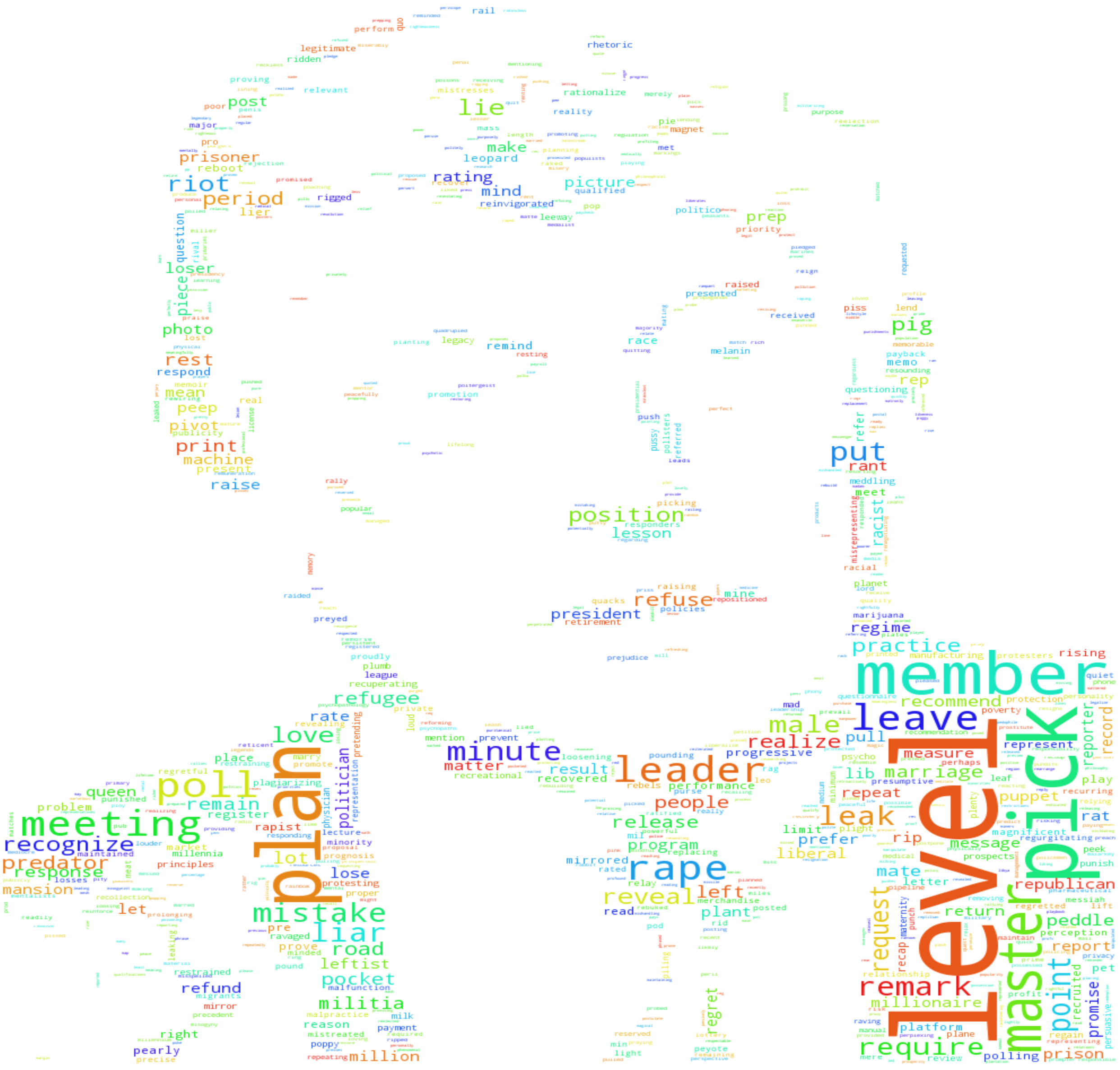


Fig 12 : Wordcloud based on trump's tweets

The word cloud showed some of the most frequent words used by trump followers in recent times. The stress on words like plan – where trump talks about having a plan of making America great , show that people did talk about Trump’s policy. Stress on words like leader, where Trump often says America lacks a leader, shows that people supported some of trump’s views.

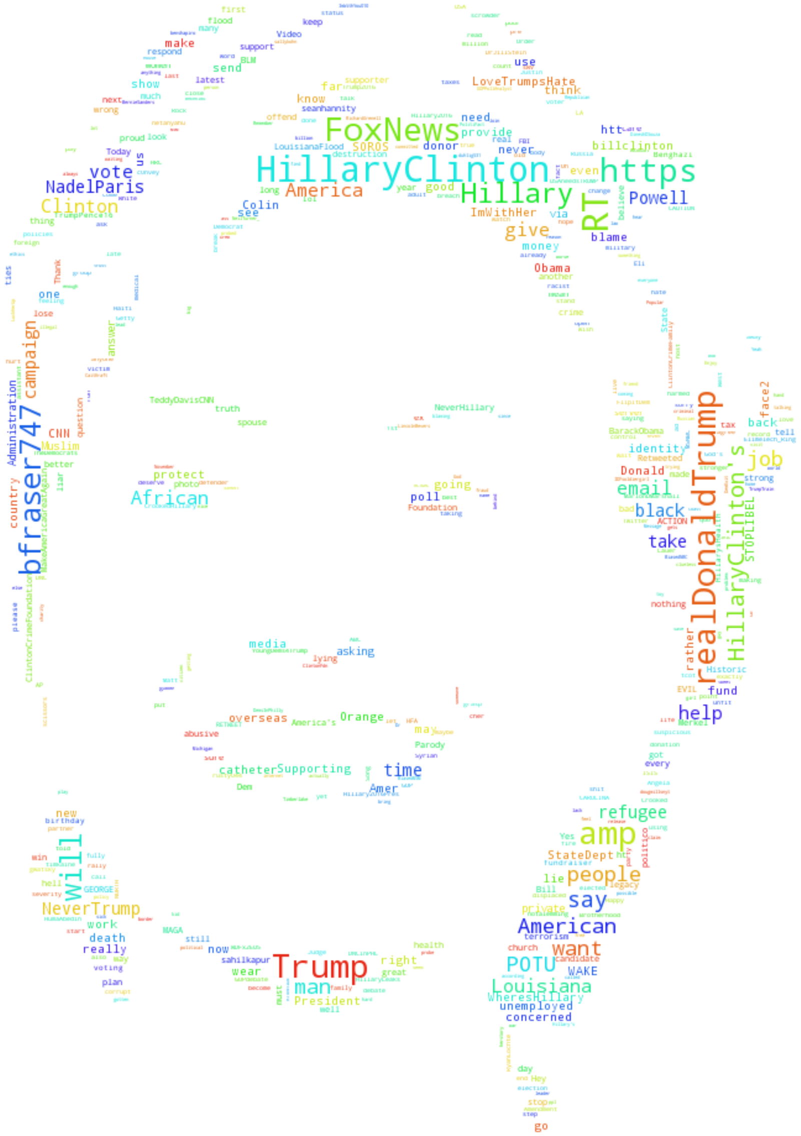


Fig 13 ; Word cloud based on Hillary’s tweets

The word cloud shows some of the common words that Hillary's followers speak in their tweets. Stress on words like emails show that people were talking about Hillary's email scam. Words like Never trump showed that people actually supported Hillary and had negative opinions about Donald Trump.

Force Directed graph

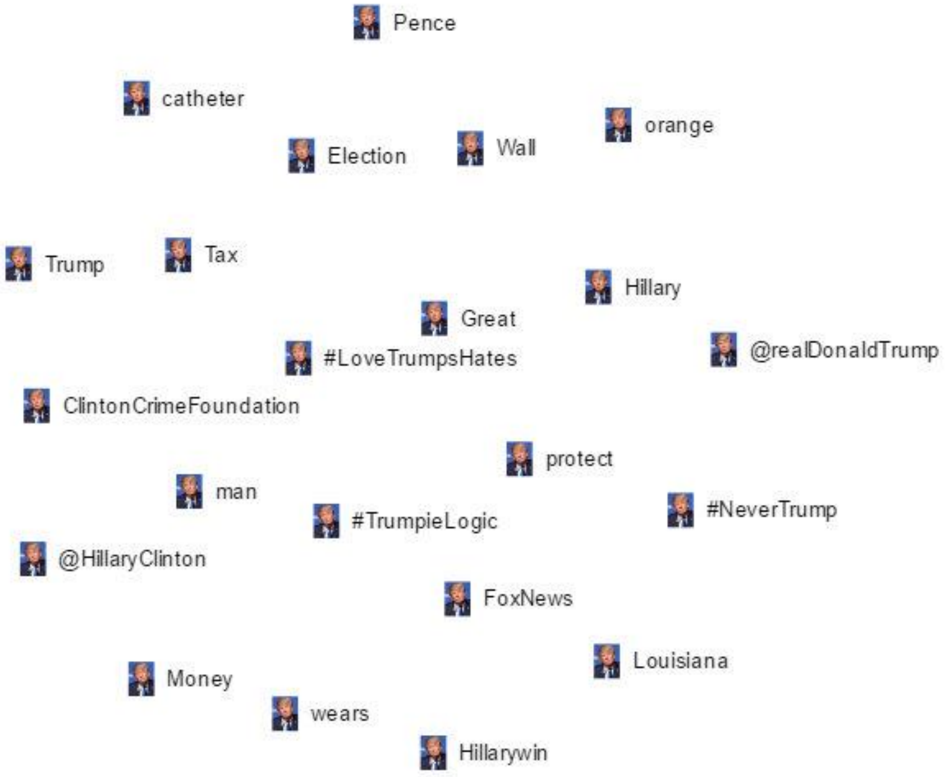


Fig 14 : Nodes of 4 grams of the most spoken words about Hillary

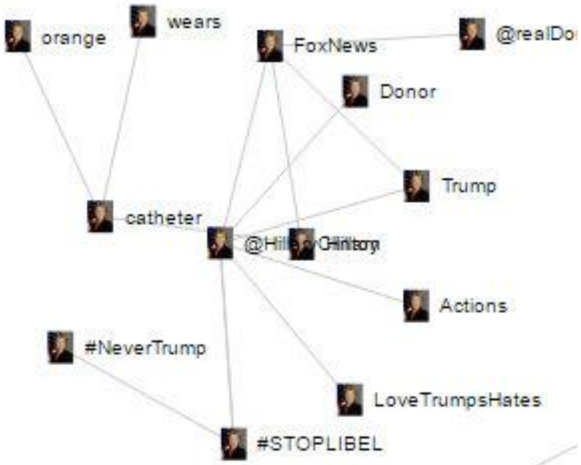


Fig 15: Force Directed Graph based on Hillary Clinton

Some of the force directed nodes like “Wears Catheter Orange” Had information which is frequently associated with Hilary. This helped us to intuitively analyze the words commonly spoken about Hillary.

Discussion and Conclusion

From the insights that we got from our visualization we were able to understand the support that Hillary Clinton and Donald Trump has all around the world. The insights and visualization provided intuitive and useful information on the twitter data.

Limitations

- Visualizing data from all over the world and comparing them would be difficult as it diverts the focus of analyzing the support for Hillary Clinton in the United States
- Our data is just a sample and it cannot be generalized to represent the entire population of the United States
- All our visualizations are not in a single platform and hence the viewer may find it hard to correlate one visualization to the other.

Caveats

- Data from social media maybe biased and prone to spammers
- Words with the intention of sarcasm can be a challenging to interpret
- Data from social media is highly unstructured and has a lot of noise.

Summary

The objective of analyzing tweets and effectively using them to describe the sentiments of people was successfully achieved based on the methods that we employed. Our data and visualizations reflected the work based on Hillary Clinton and Donald Trump. Our next objective would be to widen our scope and extract tweets county wise and perform sentimental analysis on that data. We compared and visualized Donald Trump's data against Hillary Clinton's data. This visualization gave us an overall picture about what people feel about the front runners for the presidential elections. We would like to widen our visualization by collecting both Hilary Clinton and Donald Trump data during various times of the month and expand our visualization to portray a time series visualization. With this we would be able to get an idea about the change in trend in the support for Hillary Clinton and Donald Trump.

References

- Bermingham, A., & Smeaton, A. F. (2011). On using Twitter to monitor political sentiment and predict election results
- Dorling, D., & Hennig, B. (2010). General election 2010. *Political Insight,* 1(2), 72-72.
- Robertson, S. P., Vatrapu, R. K., & Medina, R. (2010). Off the wall political discourse: Facebook use in the 2008 US presidential election. Information Polity, 15(1, 2), 11-31.
- Social networking service. (n.d.). Retrieved October 19, 2016 from https://en.wikipedia.org/wiki/Social_networking_service
- Twitter. (n.d.). Retrieved October 20, 2016 from https://en.wikipedia.org/wiki/Twitter
- Wang, H., Can, D., Kazemzadeh, A., Bar, F., & Narayanan, S. (2012, July). A system for real-time twitter sentiment analysis of 2012 us presidential election cycle. In Proceedings of the ACL 2012 System Demonstrations (pp. 115-120). Association for Computational Linguistics.

- http://comp.social.gatech.edu/papers/icwsm14.vader.hutto.pdf

Appendix

Link to code

Link to website

http://cgi.soic.indiana.edu/~arunsank/submit.html