Opinion Mining of Tweets for US Presidential Election

Subhendu Saha (sxs156132)

Nipun Agarwal (nxa150830)

Sunny Anand (sxa151231)

Abstract:

People these days take to online social networks to post their views on current topics. These topics might refer to one's personal lives, movies, politics, etc. In these project we concentrate on the political views of the people. We extract tweets from Twitter to perform a Sentiment Analysis on the presidential candidates Donald Trump and Hillary Clinton along with Bernie Sanders. We performed state wise analysis of the results and provide a visual representation of the same.

Introduction:

Online social networks such as Facebook, Twitter, Matrimonial websites, Dating websites, etc. are increasingly used by people who have access to the internet. These social networking sites enable users to publish details about themselves, connect with their friends or person of interest, publish photos, etc. One of the most important aspects of social networking sites are that people often share their views or voice their opinions on various topics. These topics might pertain to their personal lives, politics, movies, etc. Users nowadays are continuously using mobile phones to connect to social networking sites which in turn make it possible to know the user's demographics. Thus users dump a lot of data on social networking sites. These data when used ethically to extract meaningful information can be useful to various businesses, political parties, etc. One can use data mining techniques to predict trends/behavior, perform sentiment analysis, etc. on these data to extract beneficial information according to the user's demographics. This year's presidential election has been one of the most controversial and unexpected among all previous elections. The Republican Party nominee [5][6], Donald Trump went up against the Democratic Party nominee [7][8], Hillary Clinton. Though there was a popular belief that Senator Hillary Clinton would win the election by a fair margin, the outcome was rather dubious. Though Donald Trump faced strong criticism before the election, he won the election by a decent margin. We thus perform social media analytics using sentiment analysis of tweets to evaluate whether posts by related to Donald Trump, Bernie Sanders[9][10] and Hillary Clinton are either negative or positive. Based upon the same we perform our own analysis using R and Tableau for text data mining and whether our analysis replicates the outcome of the results or not. Along with machine learning and predictive modeling, we take advantage of various R packages and Tableau visualizations to analyze the sentiment of the tweets of various candidates and display the same using a format that is visually appealing. We thus map the sentiment of the tweets which gives us information of the areas where Donald Trump's comments are perceived as positive or negative. Based on the positive or negative

sentiments we try to determine the chances of Donald Trump winning the presidential election.

Background:

Many people have attempted to predict the outcome of the 2016 presidential election which is said to be extremely unpredictable because both the presidential candidates [5][7][9], Donald Trump and Hillary Clinton entered the election on back serious controversies and allegations[11]. New York Times too posted an article recently talking about how the data had failed them in the prediction of the election outcome. Even 'Google Consumer Surveys' which is rated as the one of the most accurate polling firms by 'FiveThirtyEight' in the 2012 election had recently launched its predicted election outcome with a sample size of over twenty thousand respondents across the United States could not predict the correct result. It is clear that the data used for prediction of election outcomes is of utmost importance to perform sentiment analysis as they often include sarcasm. Voters have clearly demonstrated that predictive analytics and election forecasting in particular is a nascent science. Some of these models were off by almost fifteen to twenty percent. The result from all major number crunchers after using predictive analytics was that Hillary Clinton's chances of winning was placed between seventy to ninety percent. Though predictive analysis can let us see things like never before, it can also prove to be a blunt instrument which might miss context and nuance. With all the big shot number crunchers failing to do a sentiment analysis on tweets to see if our results replicates the actual election outcome or not.

Approach:

We are using R programming language for statistical computation of the data obtained from Twitter. R is a widely used language among statisticians, data miners, etc. for data analysis as it is lightweight and easy to use. One can also visualize the data using R's software environment as well. R's popularity in recent years has increased drastically. Another advantage of using R is that it is freely available. There are various graphical front-ends available for R which makes it extremely user friendly. In terms of data mining in R, one can use several techniques such as classification, clustering, time-series analysis, etc. If needed one can write C, Java, etc. code to manipulate R objects directly which makes it easily extensible. Apart from these there has been continuous contribution on terms of packages in the R community and also includes user-submitted packages for specific functions as well. To exploit the Graphical User Interface of R we are using R Studio which is a cross platform, free and open source IDE.

For visual representation we are using Tableau. It empowers one to see and understand the data through visualization. Tableau makes analyzing of data fast and easy along with beautiful representation either using maps, charts, etc. Thus one can explore and analyze data in seconds via drag and drop options to discover trends or outliers. We use this software to produce interactive data visualization. Tableau incorporates a mapping functionality which enables us to plot latitude and longitude coordinates along with the feature of custom geo coding. Tableau also comes as a mobile application which makes it even easier to use. We get a lot more insightful and actionable information from visualizations. It also provides us with real time analytics.

We use the twitter library to extract two thousand tweets each related to Donald Trump,

Hillary Clinton and Bernie Sanders from different cities in the United States such as Washington DC, New York, San Francisco, etc. We use the longitude and latitude of these cities to extract tweets. We create a Data Frame obtained from the tweets along with the city's latitude and longitude data. We then obtain a word cloud of the words which are associated with tweets containing Donald Trump, Hillary Clinton or Bernie Sanders. We then perform reverse geocode to extract the city names from which these tweets are being generated. We analyze the tweets against positive and negative word list obtained from UIC dataset [4].

We then save the data in CSV format and move on to Tableau for visual representation of the data. We use the map representation to plot all the longitude and latitude on the map along with the respective tweets from that location. The reason for adopting this approach was to allow us to see what could have changed when we analyze the social media putting Bernie Sanders in the picture with both Hillary and Donald. The result is a great case study of the fact that indeed Bernie Sanders was the most positive political candidate.

Result & Analysis:

The experiment is broken down phase wise as shown below:

Results:

Phase I: Twitter Setup

1. Setup the credentials for the developer access key for the twiiter APP.
2. Use those credentials and save them in the current working session of R.
3. Setup the working directory for the current project.
4. Download the certificate file & Create the authenticate object for direct login into the twitter App.
5. Save this setup for the current session

Phase II: Tweets Collection

1. For each query request we would request 2000 tweets in a radius of 250 miles
2. We selected a total of 5 cities
3. We searched twitter for 3 candidates – Bernie Sanders, Donald Trump, Hillary Clinton
4. Collected the tweets and collected features from these tweets and form the dataset.

Phase III: Creation of Corpus & Word Cloud

1. Create a corpus by removing stop words, punctuations and White space.
2. Create the word cloud for each candidate

For Hillary Clinton

For Donald Trump

For Bernie Sanders

Phase IV: Geolocation of Tweets

1. Create a corpus by removing stop words, punctuations and White space.
2. From each of the data rows find the address, city, stzip, zipcode, state

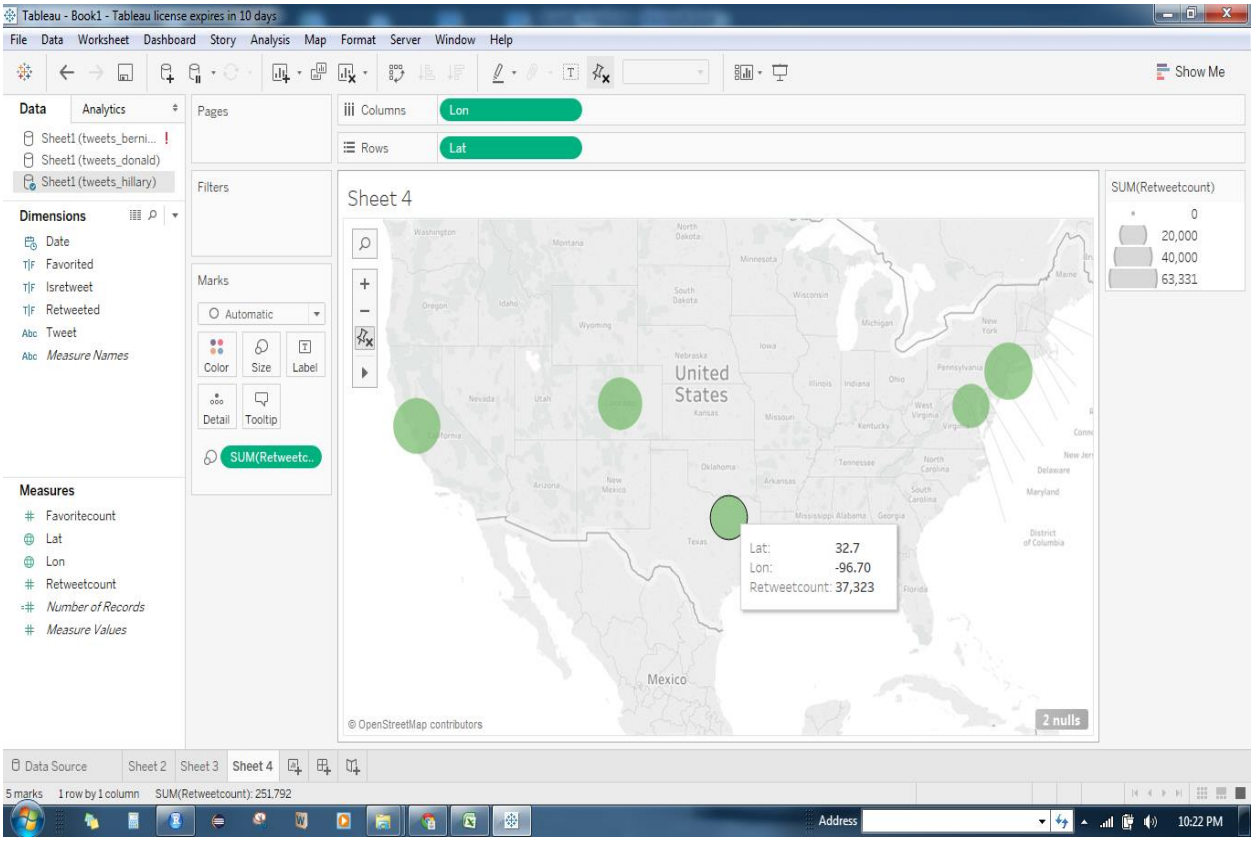
Phase V: Sentiment Analysis of the Tweets location based

1. Get the list of positive and negative words so we can compare each word
2. Calculate the score based on the index position of each word in the tweet
3. Determine the sentiment scores based on these positive and negative words in the tweets

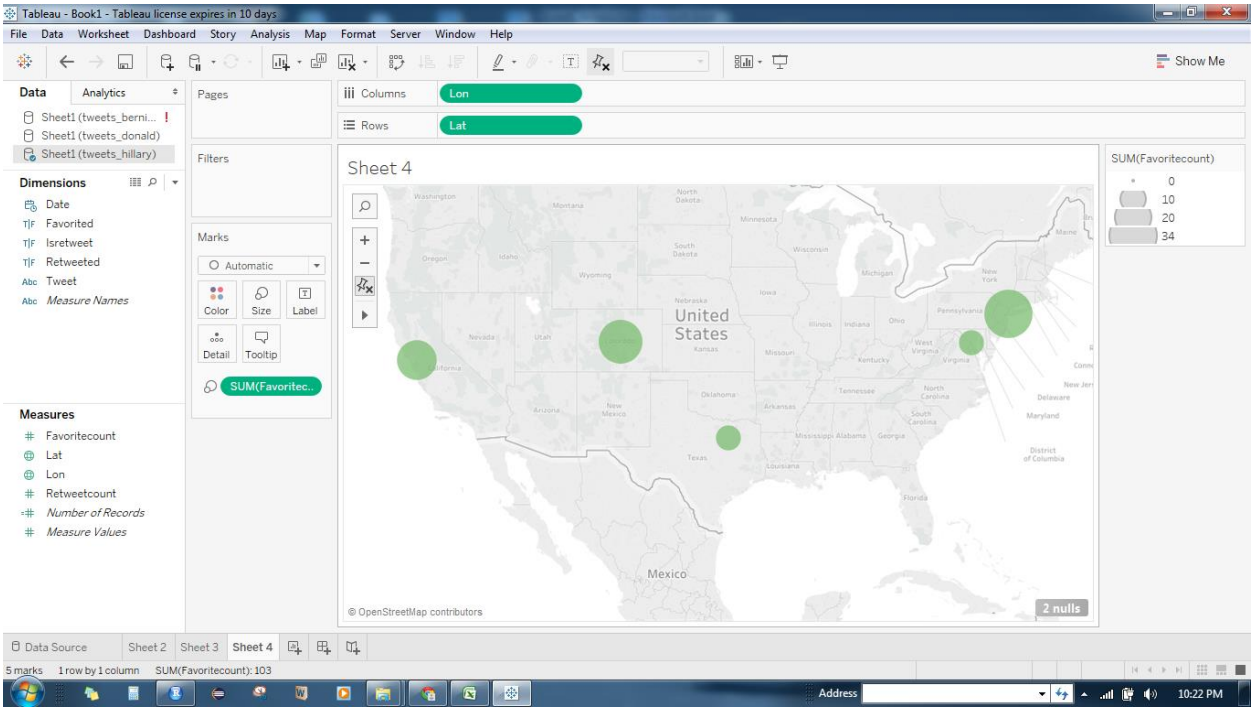
Phase VI: Visualization using Tableau

Tweets Distribution for Hillary Clinton

a.) No of Retweets

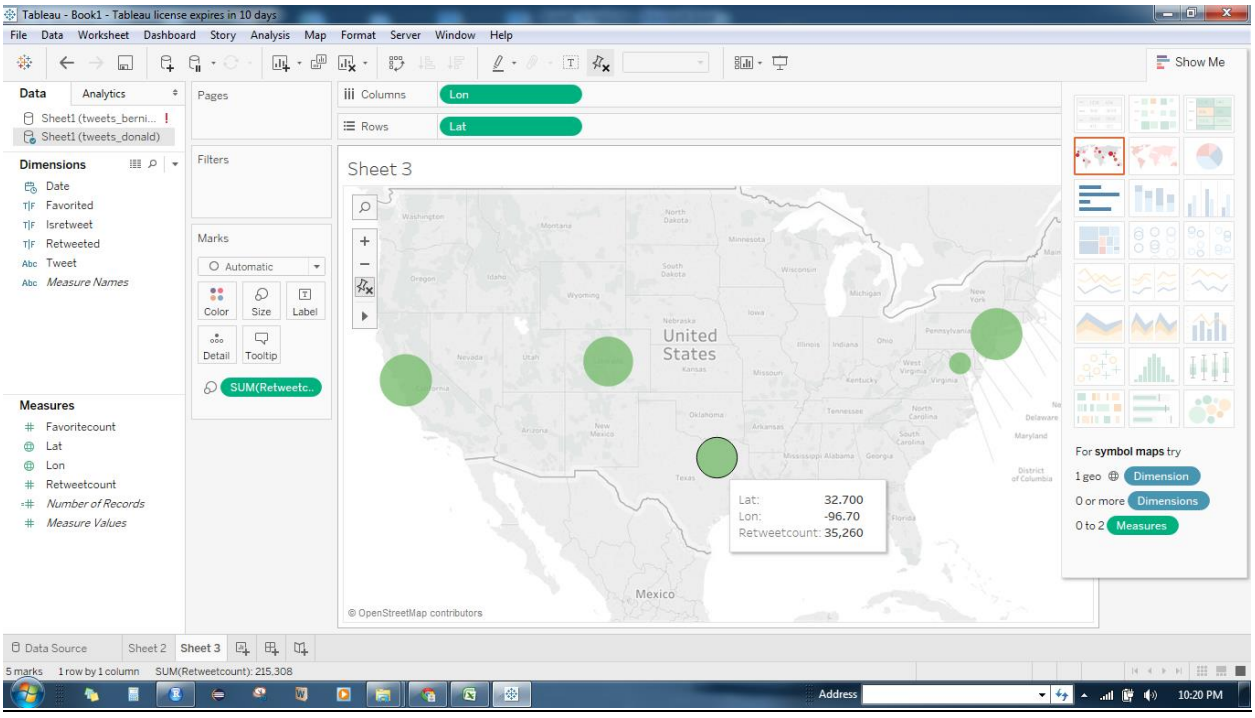


b.) No of Favorites

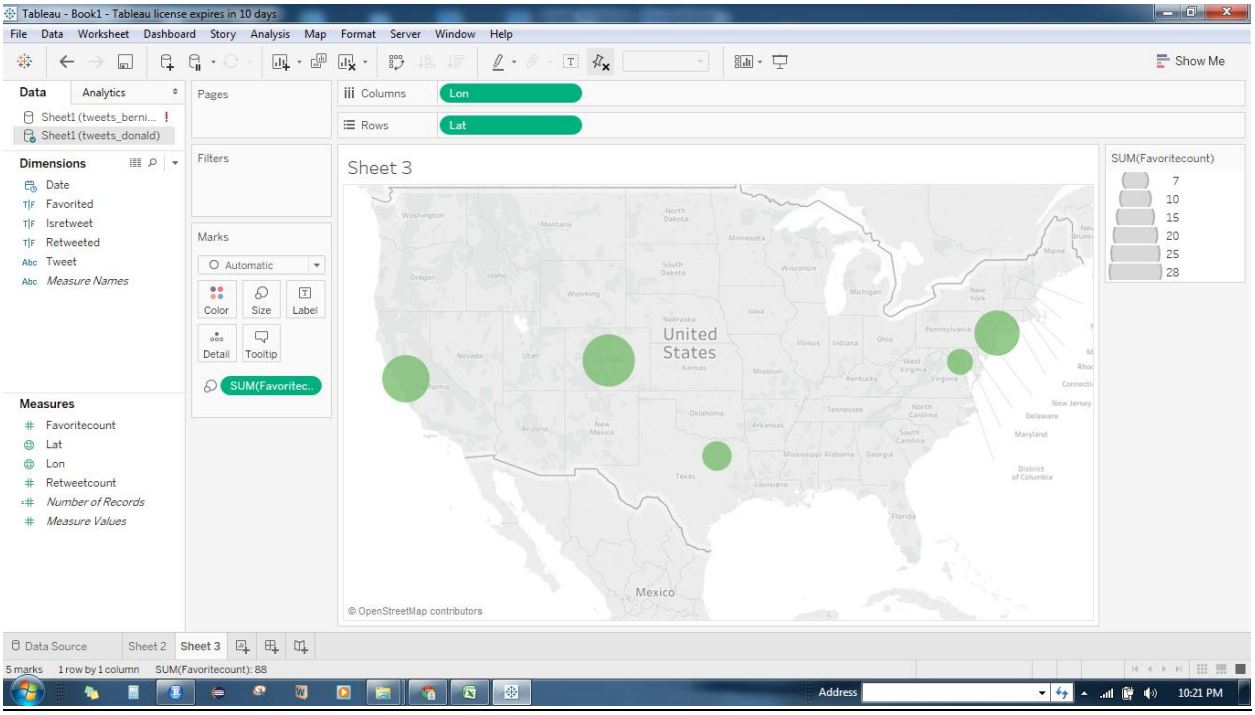


Tweets Distribution (Donald Trump)

a.) No of Retweets

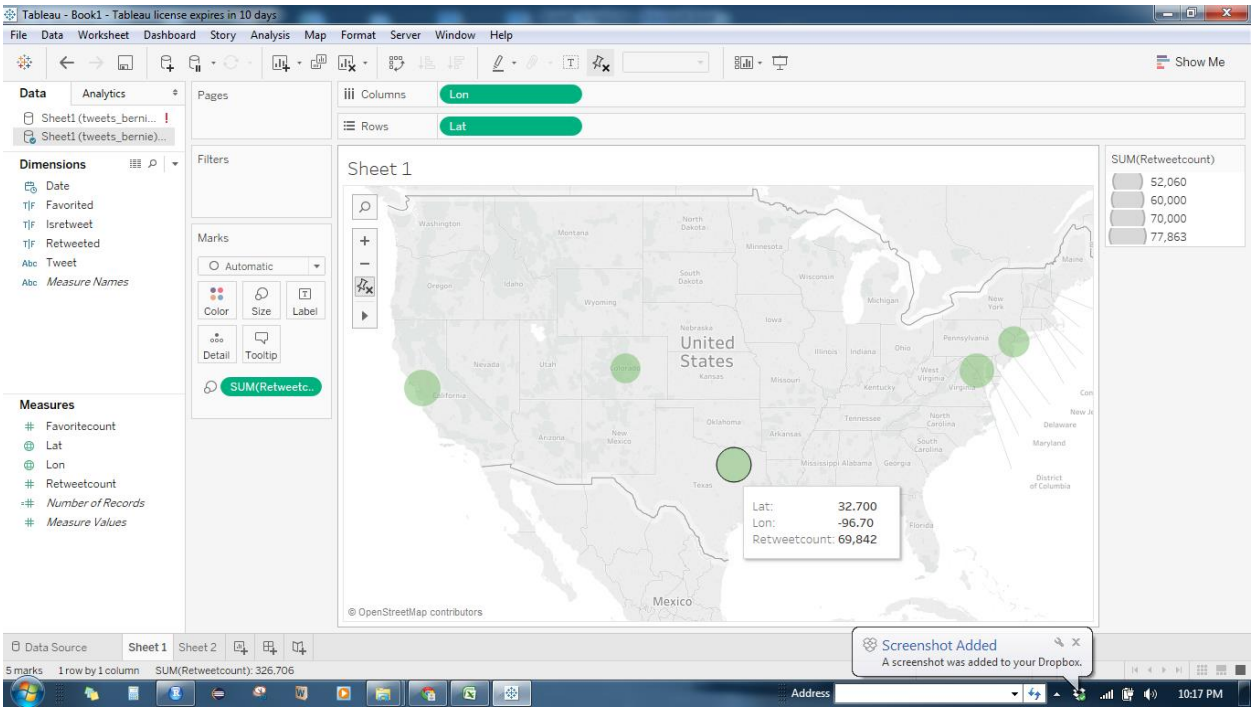


b.) No of Favorites

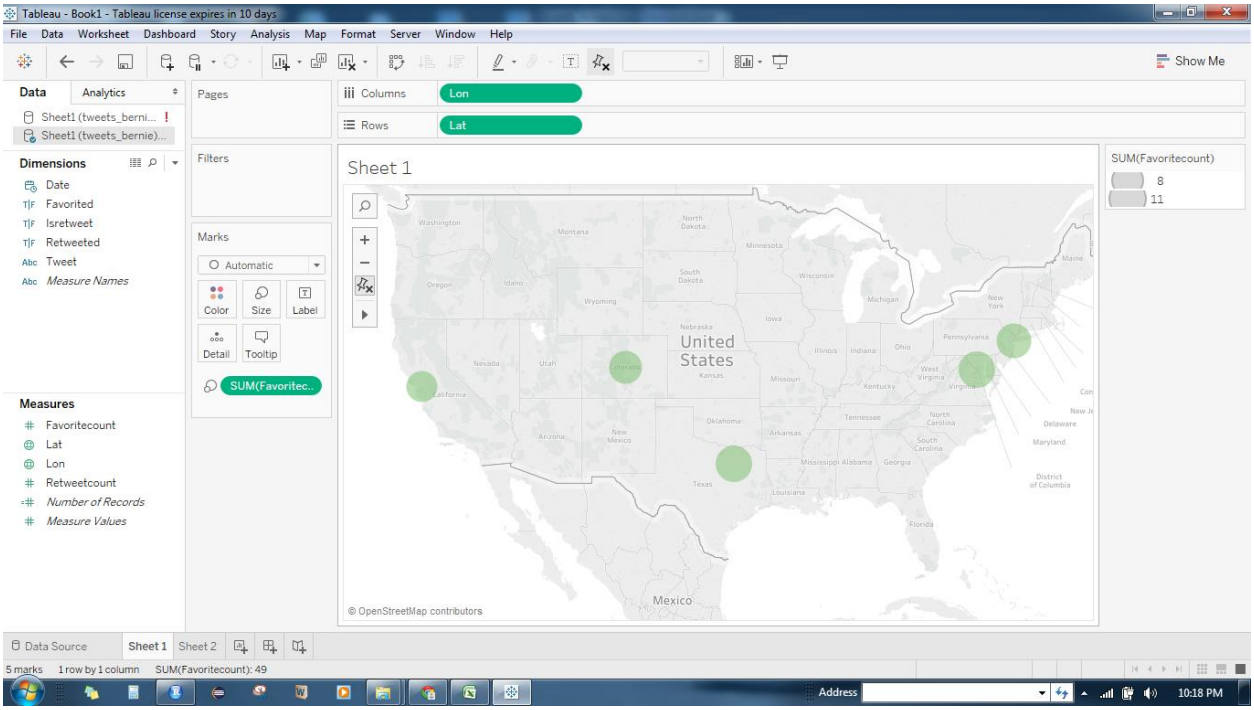


Tweets Distribution (Bernie Sanders)

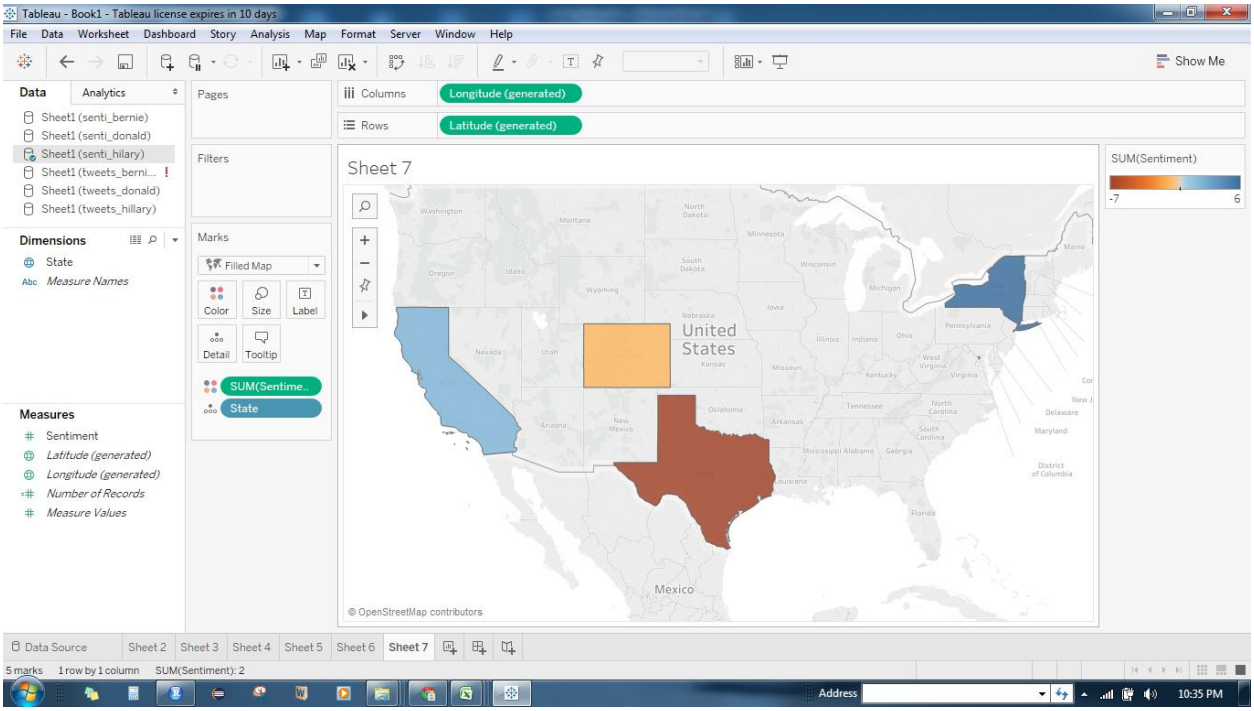
a.) No of Retweets



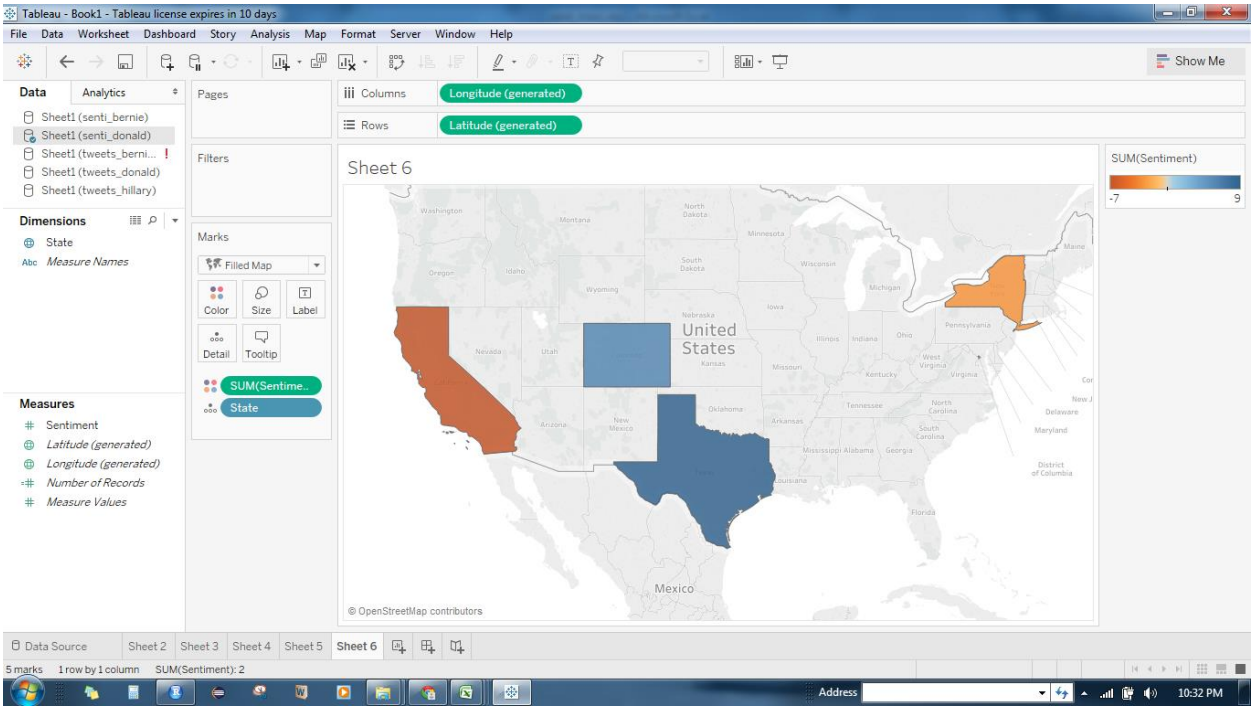
b.) No of Favorites



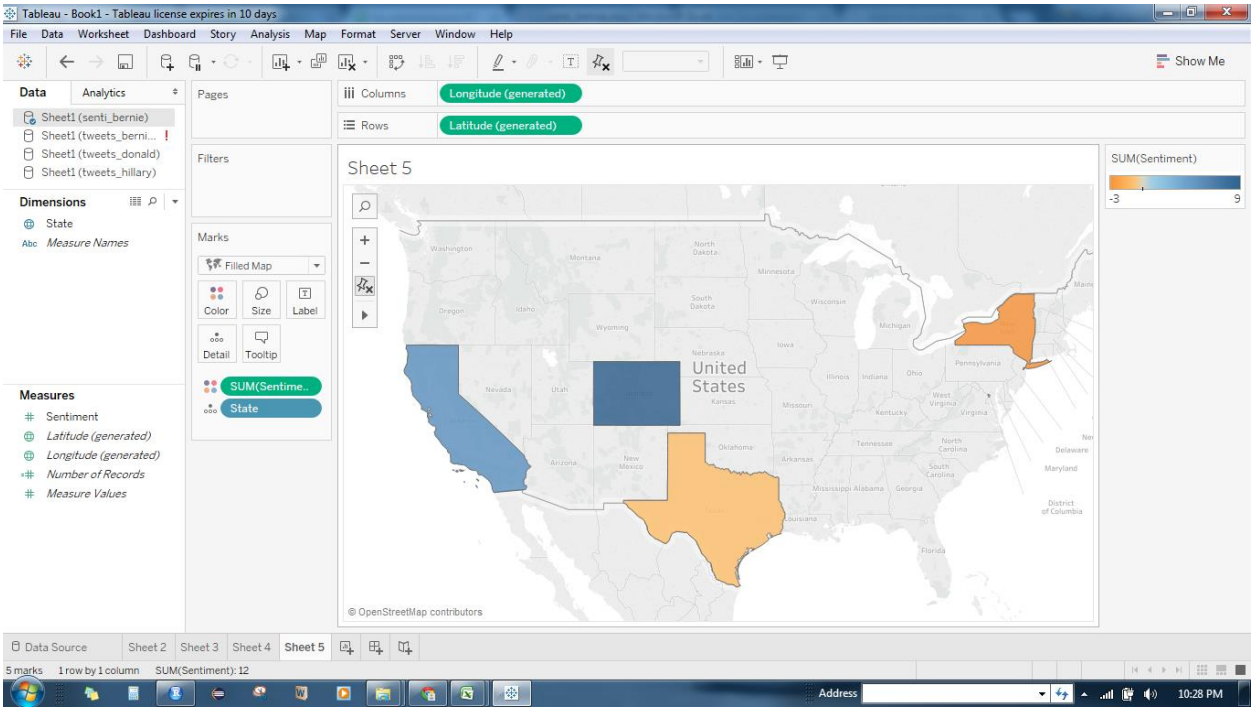
Sentiment of Sample Tweets That Have Hillary Clinton in Them



Sentiment of Sample Tweets That Have Donald Trump in Them



Sentiment of Sample Tweets That Have Bernie Sanders in Them



Analysis of Result:

The results from the sentiment analysis clearly show Bernie Sanders to be a strong candidate against Donald Trump. Bernie Sanders seems to take the state CO(Colorado) where as Hillary Clinton seems to be a weak candidate there when pitted against Donald Trump. With Sanders it seems quite clear that Democrats had a much better chance winning CO, NY and CA with Trump only wining TX.

Future Work:

In this project we just attempted to gather sentiment analysis from tweets. In future we would incorporate social media networks such as Facebook, Instagram, etc. We will also try and analyze how swing states will affect the elections. We will expand our sentiment analysis in states beyond the United States of America. We will also try and extract the sensitive topics which are driving the election so that the candidates can emphasize on those topics. Since all these statistics will be demographic, a candidate will know precisely in which states he or she should put in more efforts to gather votes.

Conclusion:

According to our in depth analysis on the twitter data we found that Bernie Sanders had the most positive response on his tweets. Donald Trump's tweets were the second one to follow with Hillary gathering the most negative response. We further conclude that though Bernie Sanders would have been an ideal candidate, Donald Trump too is not that far behind. He does not need to change the manner in which he tweets as most of his tweets are perceived as positive by the voters.

However the election polls showed Hillary to be the clear winner, it seems that the people decided to choose the less negative person out of the option available to them. We strongly believe that it was these sentiments which were not captured in poll but the small survey conducted by poll exit companies which were not a true representation of the American voters and their feeling showed a not so accurate result. Our experiment to analyze the most ideal candidate and it was no surprise it was Bernie Sanders proves that social media networks are playing a vital role in creating and influencing opinions among people on such important aspects. For visualization we used the map plot distribution of tweets across the United States for a quick understanding of the sentiments of tweets and the widespread opinion of people about a candidate in that particular region.

References:

[1]https://analytics.googleblog.com/2016/09/google-consumer-surveys-launches-weekly.html

[2]https://en.wikipedia.org/wiki/R_(programming_language)

[3]https://en.wikipedia.org/wiki/Tableau_Software

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

[5] https://www.donaldjtrump.com/

[6]https://en.wikipedia.org/wiki/Donald_Trump

[7]https://www.hillaryclinton.com/

[8]https://en.wikipedia.org/wiki/Hillary_Clinton_presidential_campaign,_2016

[9]https://go.berniesanders.com/page/content/splash?source=homepage

[10]https://en.wikipedia.org/wiki/Bernie_Sanders_presidential_campaign,_2016

[11]http://www.theatlantic.com/politics/archive/2016/11/tracking-the-clinton-controversies-from-whitewater-to-benghazi/396182