Trump Tweets Analysis

Project Report

Group 13

Chaitanya Gokhale

Vijaya Vasavi Seenivasan

Vamsinadha Reddy Mallavaram

**Problem Statement and Background:**

Donald J. Trump became the 45th President of the United States on January 20, 2017.Since the inauguration, President Trump is actively initiating new policies and conversations, which generate active conversations in the social media. As Data Scientists, we want to take advantage of this opportunity by using text-mining approaches to conduct social media analytics

**Project Overview:**

This project is to analyze people’s sentiment and topics about the new administration. We will use Twitter API to collect tweets about Trump. Then we will conduct sentiment analysis to measure how positive or negative the collected tweets are, which can be an indirect measure of Trump’s approval. Next, to see what kinds of topics are discussed related to the new president, we will create word clouds and conduct topic modeling on the collected tweets. To see the geographic variation in opinions, we will collect tweets from 5different states and conduct the aforementioned three analyses. Finally, please conclude the project by describing the insights you gained based on the conducted analyses.

**Data Collection:**

As our analysis needs to be done around Donald J. Trump, the initial analysis need tweets that consists/relate anything to *Trump*, which are collected by using Twitter Streaming API’s. The package we used to handle the data collection in python is Tweepy and the target keyword as “trump”.

Since our experiment design is to also test the geographic variation in the opinions, we have collected tweets from other states i.e., Texas, California, New York, New Jersey and Florida. To achieve this, the tweets were filtered on user location & classified when desired State is found. These results were stored into 5 different files for further analysis.

**Data Cleaning:**

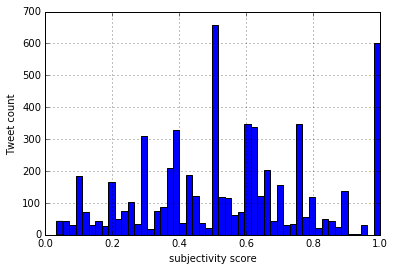
One of the main challenges we faced during this project is data cleaning. While twitter allows user to input several special characters along with usual punctuations, few other characters/Letters refer to certain actions on twitter i.e., @userName to refer a person in the tweet, RT for Re-Tweet. Such characters and strings that add no sense to the analysis needs to be removed from our corpus. Also, we should throw out any web links in our data which are not appropriate for text analysis.

**Note:** Though (‘) is a punctuation, it is often used as diacritical mark in the text than it is used as a separate special character. And replacing this with a space will make the word incomprehensible. So we have replaced it with nothing to preserve the meaning of the word. i.e., let’s => lets, couldn’t => couldn’t, who’s => whos etc.,.

**Sentiment Analysis:**

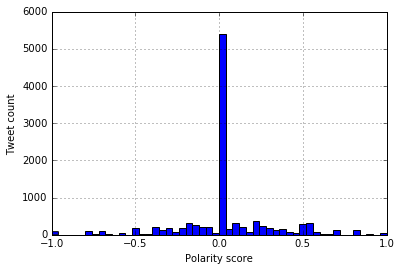
To measure the sentiment of the people who tweeted about trump, we calculated the subjectivity and polarity scores of the whole corpus & the summarized scores are as displayed below:

*a.) subjectivity score:*



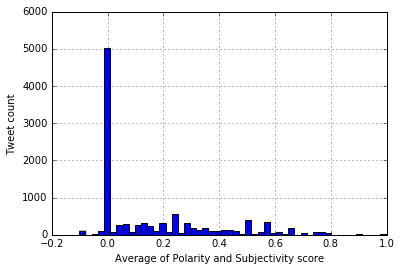
The subjectivity score distribution is showing that – though good portion of our corpus is not really subjective towards any topic/opinion, most of it have medium to very strong opinions pushing the Avg.weight of subjectivity higher, thus revealing the fact that majority tweets have specific opinion than a general/common opinion.

*b.) Polarity Score:*



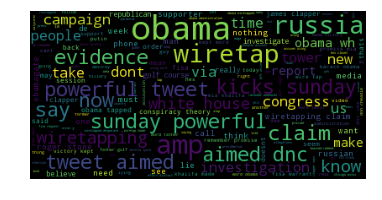
The polarity distribution of the corpus is showing that most of the tweets take a neutral stand, though being subjective.

*c.) Average of Subjectivity & Polarity:*



By plotting the average of subjectivity & polarity we can visualize that most of the users are taking a neutral stand w.r.t their emotion while not making a specific opinion. On the other hand, those who have positive opinions are also specific in expressing their opinions.

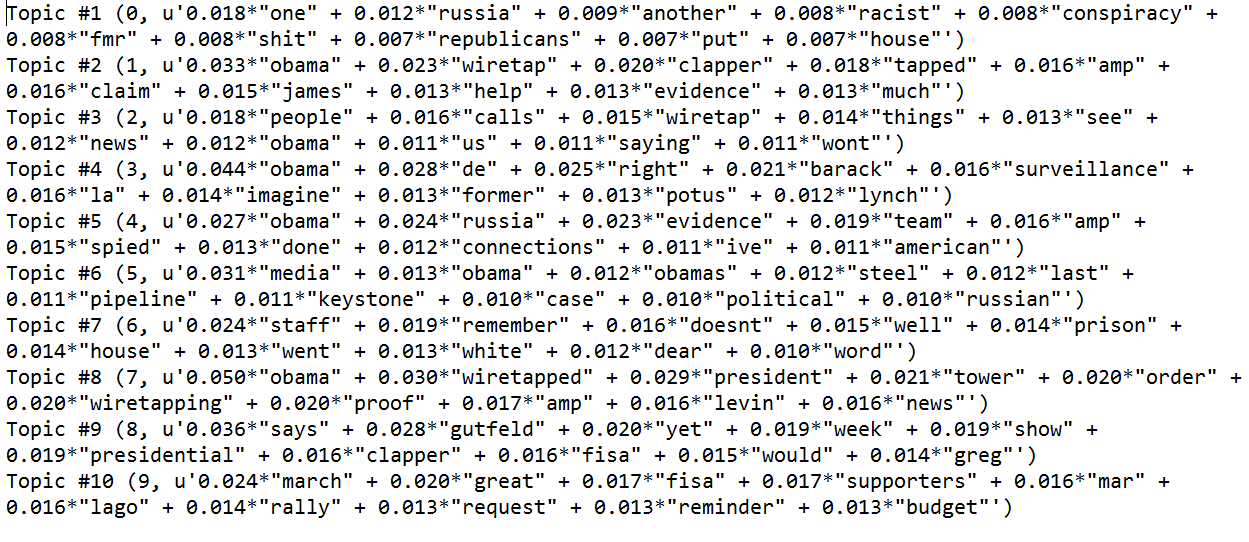
**Word Cloud for 10k Tweets:**

****

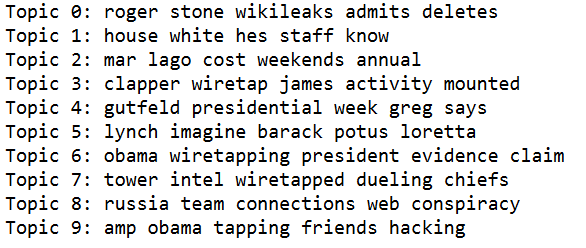
From the word cloud, it looks like people are more speaking on the subject: Trump's **claim**, **Obama** had **wiretapped** him during the **campaign.**

**Topic Modeling:**

a.) LDA:

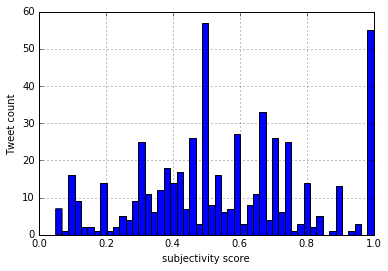


*b.) NMF:*

**

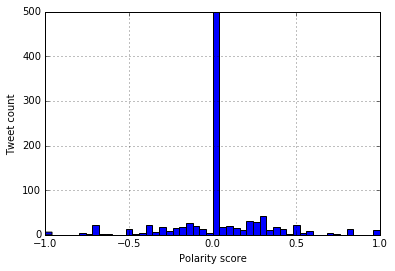
**Texas State**

*a.) subjectivity score:*



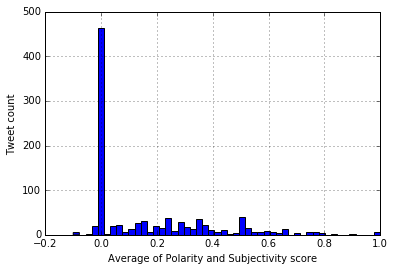
The above plot is showing that we have a fair distribution of subjectivity scores, which states that users from Texas states are making opinions but not very strong opinions.

*b.) Polarity Score:*

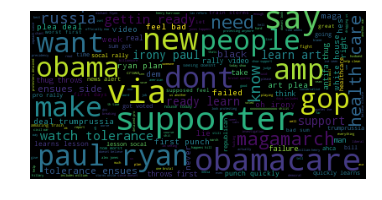


The polarity distribution of the corpus is showing that most of the tweets take a neutral stand.

*c.) Average of Subjectivity & Polarity:*



**Word Cloud:**

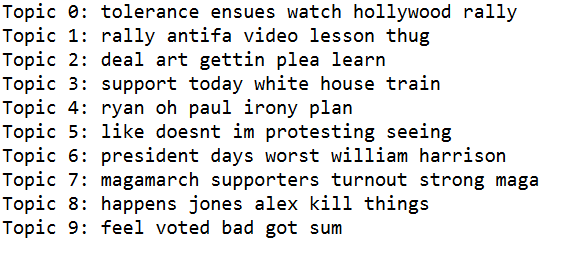
****

**Topic Modeling:**

a.) LDA:

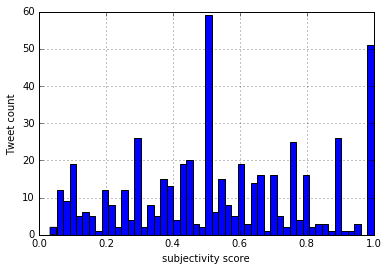


*b.) NMF:*

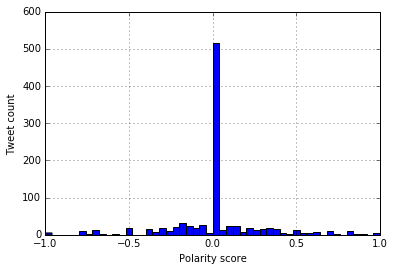
**

**California State**

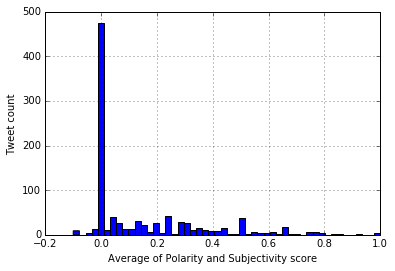
*a.) subjectivity score:*



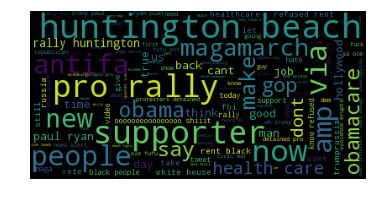
*b.) Polarity Score:*



*c.) Average of Subjectivity & Polarity:*



**Word Cloud:**

****

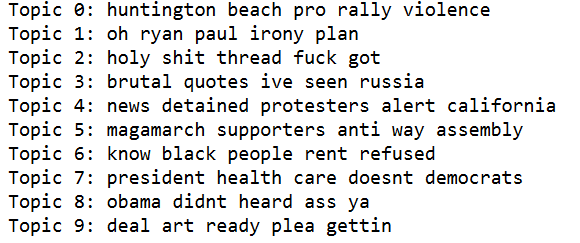
From the word cloud, we can summarize that people tweeted more about the **Trump’s** **Pro** **Rally** in **Huntington** **beach.**

**Topic Modeling:**

a.) LDA:

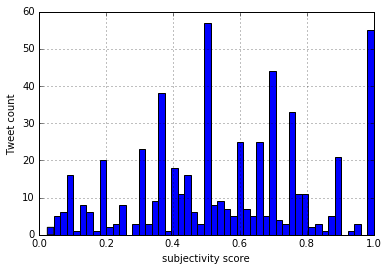


*b.) NMF:*

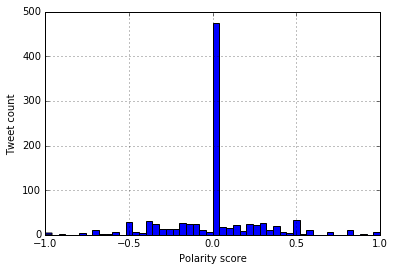
**

**New York State**

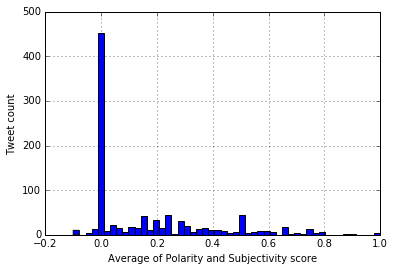
*a.) subjectivity score:*



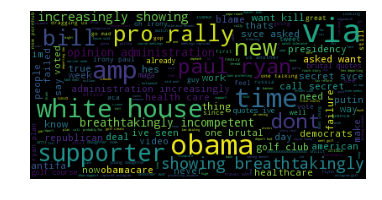
*b.) Polarity Score:*



*c.) Average of Subjectivity & Polarity*



**Word Cloud:**

****

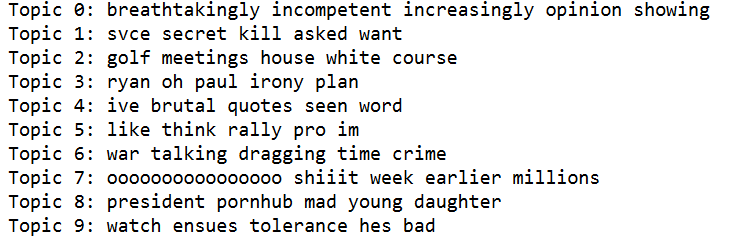
From the New york’s word cloud: While people are talking about the pro rally, they look like tweeting more about the Obama care healthcare reform bill and it’s failure.

**Topic Modeling:**

a.) LDA:

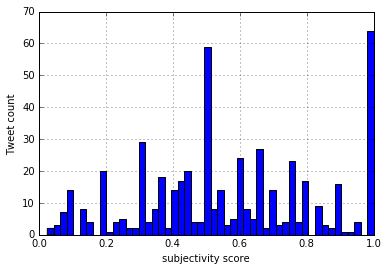


*b.) NMF:*

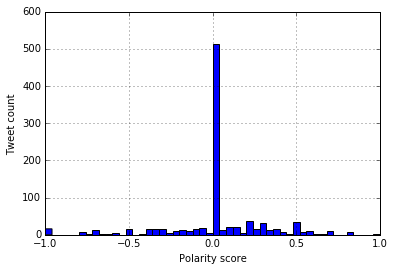
**

**Florida State**

*a.) subjectivity score:*

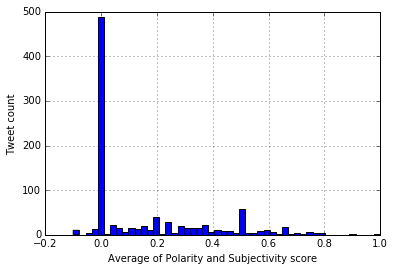


*b.) Polarity Score:*

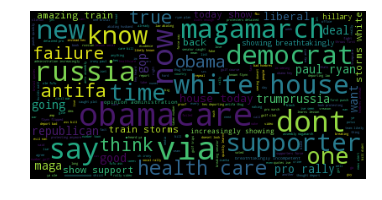


The polarity distribution of tweets from Florida shows some strong negative emotions tweeted by the people.

*c.) Average of Subjectivity & Polarity:*



**Word Cloud:**

****

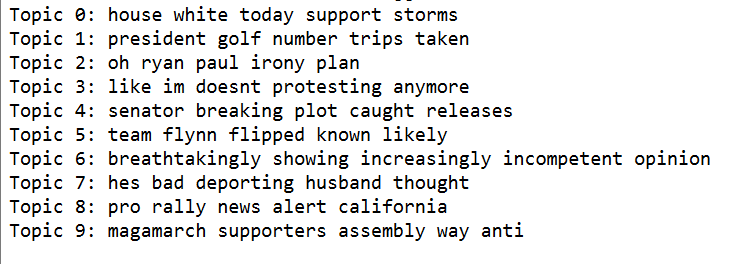
Tweets by the users from Florida have more references to the word “Russia”, “antifa”, “magamarch”**.**

**Topic Modeling:**

a.) LDA:

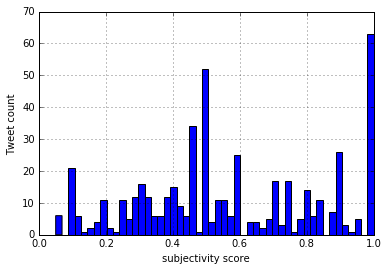


*b.) NMF:*

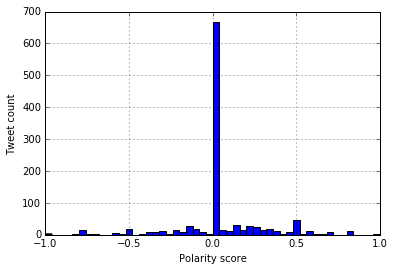
**

**Arizona State**

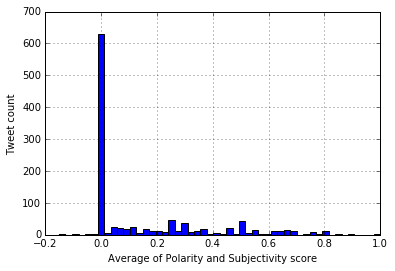
*a.) subjectivity score:*



*b.) Polarity Score:*



*c.) Average of Subjectivity & Polarity:*



**Word Cloud:**

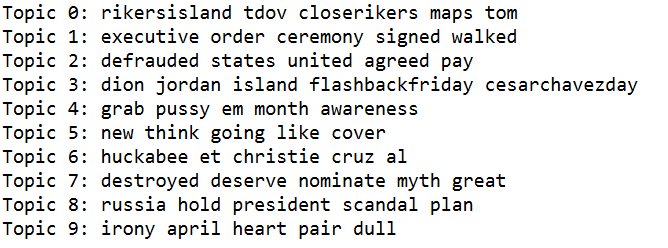
****

**Topic Modeling:**

a.) LDA:



*b.) NMF:*

**

**Insights drawn from analysis:**

1.) Subjectivity score is highest in the state of Texas, by which we can say that the users in the Texas are being more particular in expressing their opinions.

2.) As polarity scores are more positive in Florida and Arizona we can cay that users from these states are favorable to trump. Which can also be backed by the fact, he has won in both these states.

3.) People from Arizona are paying more tribute to the musical icon “selena” than in any other state, even more than her born state Texas.

4.) “General opinion in the New York about Trump’s administration is increasingly incompetent” – this can be concluded from the LDA and NMF topic models.

5.) New York & Texas has more tweets in the range of -0.5 to +0.5 that are not falling on 0.0. This could possibly mean that users from these states are more comfortable in showing their emotions (both positive & negative) on twitter.

6.) California tweeted more about the **Trump’s** **Pro** **Rally** that happened in **Huntington** **beach,**CA. But the other states were surprisingly silent about this Rally.

7.) Arizona is talking more about **rikers island**, after the announcement by the city mayor that it is going to get be closed. And the main top keywords are ‘**rickersisland’, ‘rickersisland google’, ‘closerickers’, ‘google maps’**. From this, we could infer that most of the users from Arizona doesn’t know about the location of this jail or the rickers lisland.