**How can Trump win?  --- Data analysis on presidential-election-related tweets**

DS501 Team 10

Phaterpekar, Ankit Manohar; Yi, Lingyun; Su, Yupeng; Guram, Jui Saba; Cui, Xiaoting

Table of Contents

[INTRODUCTION](file:///D:\Downloads\Sample%20Report.docx#_Toc442751631) 3

[DATA GATHERING 3](file:///D:\Downloads\Sample%20Report.docx#_Toc442751632)

[DATA STORAGE AND VISUALIZATION 5](file:///D:\Downloads\Sample%20Report.docx#_Toc442751633)

[DATA ANALYSIS](file:///D:\Downloads\Sample%20Report.docx#_Toc442751635) 5

[MUTUTAL FRIENDS AND FOLLOWERS](file:///D:\Downloads\Sample%20Report.docx#_Toc442751636) 7

[BUSINESS CASE](file:///D:\Downloads\Sample%20Report.docx#_Toc442751637) 9

[CONCLUSION AND LIMITATIONS 10](file:///D:\Downloads\Sample%20Report.docx#_Toc442751638)

**Introduction:**

Every second, on average, around 6,000 tweets are tweeted on Twitter. In total, twitter has [310 million monthly active users](http://techcrunch.com/2016/04/26/twitter-reports-mixed-q1-on-sales-of-595m-ep-of-0-15-and-a-sluggish-310m-maus/), while US has 65 million active user per month. Twitter has become a popular online micro-blogging service that highly influence the US people’s opinion on things, especially politics. The 2008 US Presidential Election was the first to be won on the Internet. From then, internet, as an election tool, starts to impact the election campaigns not only in US but also in other countries. Politicians become much more active on Twitter. With the fact that [83% of the world’s leaders](http://www.prnewswire.com/news-releases/world-leaders-on-twitter--adoption-stagnates-even-as-follower-base-explodes-300208802.html) are on Twitter, people no longer reply on traditional media, like TV, Newspaper, to get the updates from their president candidates or leader. By posting the tweets, politicians could pass their opinions to public to advance their personal brand. Public use Twitter to comment on the political topics and, moreover, they could interact with Politicians directly online. A vast online political sphere gets built up.

Based on the latest state and national polls, the New York Time’s Upshot election model predicts there is only 25% chance Donald Trump of winning the presidency. To continue the campaign, how could Donald J. Trump win more votes? The first step to answer that question is to understand what people think and then adjust campaign strategy accordingly. The big data on twitter is the best source to find out the political issue people care the most during this presidential election. In this study, with the analysis on the twitter data, we are aiming to find out the public focus and common desire by looking for the words showing up with highest frequency, state trends and the most influential users or groups within 2016 election related tweets. The analysis results will be a good base to come up with optimal campaign strategy for Donald Trump.

**1.Data Gathering:**

In order to evaluate the current sentiments surrounding the 2016 Presidential Elections, the Twitter streaming API was used to sample tweets in real-time.

Here our objective to analyses the tweets is to help the campaign team of Donald trump to increase his support and give the campaigning team some base to work their strategy on. We have collected the tweets in Morning, afternoon and night to avoid any bias from time difference between states.

Tweets containing the key-words mentioned below within the text body.

Key Words:

* 'Trump2016'
* 'Hillary2016', '
* #imwithher',
* '#makeamericagreatagain',
* 'election2016',
* 'Donald Trump',
* 'Hillary Clinton',
* 'presidentialelections2016',
* 'donald trump'
* 'hillary clinton' “

**Some of the questions we have tried to answer are**

1. **What political issues people talk about the most?**
2. **What tweets influence people the most?**
3. **Which tweeter user is tweeting the most for Donald?**
4. **What are the sentiments (support,hate) related to the tweets for Donald?**
5. **From what part of the world are people tweeting the most?**
6. **Do people tweet about Donald more or Hillary?**

**Results:** These records were collected and stored in a JSON file (tweets.json).

A total of 127,367 tweets were obtained as a result of this process.

By observing these tweets we found out that we could see the geographical location of the tweeter user ,the reweet count in it and also what is the retweet.

**Data Storage and Visualization:**

The JSON file was then imported into MongoDB. For this the mongo import functionality was used. MongoDB was the preferred choice as the document-oriented database allows for fast storage and retrieval. In addition it allows to summarize the results through its aggregation framework. In addition Elasticsearch was used for to allow a search engine based functionality on the dataset. Kibana Plugin was used with Elasticsearch to provide data visualizations capabilities on top of the content indexed on the Elasticsearch cluster.

**Data Analysis:**

**Frequency Analysis**

We performed frequency-based analysis on this collection of tweets. We determined the most commonly occurring words in the collection of tweets. To achieve this we used the NLTK. corpus library along with the use of regular expressions to filter the stop words, some common emoticons and punctuations.

**a) Top 30 words**

The results for the top 30 words were as follows:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **TOP 30 Words** | | | | **TOP 30 Words** | | | | |
| **Nouns** | **Count** | **Adjctives** | **Count** | **Nouns** | **Count** | | **Adjctives** | **Count** |
| American | 1255 | Prison | 8546 | Republican | 538 | birther | | 1113 |
| safe | 1177 | president | 2297 | political | 533 | world | | 1090 |
| much | 1108 | Ohio | 1955 | next | 518 | donald | | 1021 |
| Safe | 1003 | vote | 1890 | many | 501 | need | | 1002 |
| new | 999 | Black | 1870 | presidential | 495 | Emmys | | 999 |
| foreign | 852 | BREAKING | 1655 | deplorable | 482 | Rick | | 986 |
| main | 850 | Again | 1566 | wrong | 477 | Poppe | | 980 |
| black | 732 | everything | 1508 | great | 475 | racist | | 971 |
| last | 716 | pregnancy | 1502 | Haitian | 459 | election | | 964 |
| continued | 715 | voting | 1482 | old | 453 | charity | | 954 |
| good | 676 | game | 1425 | unprepared | 398 | candidate | | 953 |
| domestic | 645 | Thank | 1423 | free | 397 | speech | | 943 |
| former | 580 | ground | 1409 | false | 393 | support | | 943 |
| real | 561 | ISIS | 1242 | white | 383 | heard | | 922 |
| bad | 557 | video | 1171 | African | 373 | Emmy | | 914 |

Table 1.1 Top 30 words

**Results:** This table of the most common words (top 30)

Answers our question of **What political issues people care the most?**

By observing the table we see that people care about safety,Haitian,foreign policies and races

**b) Most popular tweets**



Table 1.2 Most popular tweets

Next, we tabulated the most popular tweet, i.e. the tweets with the largest number of retweet counts.

**Results:** By observing this table we come to know which tweets are influencing the people the most? It will give an insight regarding what people what is the most popular topic or candidate people are talking about.If we know the most popular topic ,Donalds campaign team can focus on those topics.

**C)** We also tabulated the most popular tweet entities – hashtags and user mentions occurring in our dataset. This were obtained used the MongoDB aggregation framework and Elasticsearch, Kibana. These results were as follows:

|  |  |
| --- | --- |
| **Top 10 hashtags** | **Count** |
| trump2016 | 10767 |
| imwithher | 8202 |
| maga | 4480 |
| trumptrain | 4067 |
| emmys | 3552 |
| makeamericagreatagain | 3177 |
| trump | 2403 |
| hillary | 1824 |
| hillaryshealth | 1536 |
| election2016 | 1506 |

|  |  |
| --- | --- |
| **Top 10 most mentioned** | **Count** |
| realdonaldtrump | 5706 |
| hillaryclinton | 5042 |
| blackpplvines | 1935 |
| kellyannepolls | 1569 |
| stylishrentals | 1564 |
| usafortrump2016 | 1518 |
| darren32895836 | 1484 |
| berniesanders | 1329 |
| timrunshismouth | 1313 |
| youtube | 1306 |

Table 1.3 Top 10 hashtags Table 1.4 Top 10 most mentioned

**Results**: We observe that Donald trump was talked about at the Emmys award.Also people are talking about Kellyanne who’s Donalds campaign manager.

**Users who are friends and followers of the Donald Trump campaign.**

Since we are interested in Donald trumps tweets we found the friends and followers from official trump campaign account since Donald trumps twitter account didn’t have many friends.

As we are interested in the 2016 election. In order to find out some information about the two candidate, Hilary Clinton and Donald Trump，we choose USAforTrump2016 as the screen name to get people who are in interested in Trump. This user has 309 friends and 49082 followers. 20 friends and followers of that user are listed in two tables. The first table is the friends table and the second table is the followers table.

**Table 3.1: Example Friends Table 3.2: Example Followers**

**of @USAforTrump2016 of @USAforTrump2016**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ID | SCREEN NAME |  | ID | SCREEN NAME |
| 261014695 | LGlick1 |  | 210204689 | TylerHinson35 |
| 17469289 | nypost |  | 348271154 | REALbigALMorgan |
| 16799621 | OMAROSA |  | 4203625033 | happynescats |
| 80082384 | RobertJohnDavi |  | 99800543 | PauleMontal |
| 56561449 | jessebwatters |  | 488060497 | DragonForce\_One |
| 19211550 | gatewaypundit |  | 3153694671 | Shelby14\_02 |
| 3981225735 | TrumpStudents |  | 1618628058 | LaneMartinsen |
| 3118705906 | joshgremillion |  | 778199093734551552 | w3bulbul |
| 202814670 | MAGA\_dot\_Media |  | 2523754471 | IhateWhiteknigh |
| 28628850 | greggutfeld |  | 778274614011162624 | DarrellAndrew12 |
| 16163639 | NolteNC |  | 4813614554 | eggs\_and\_Laken |
| 2465207048 | BennettJonah |  | 1649293314 | gbielan |
| 2645181774 | NYPDCT |  | 761763406919585796 | conservnews |
| 292777349 | pzf |  | 778468907514368000 | hillarysmells |
| 3327736274 | AngelTilaLove |  | 1967394264 | GOTOREALTORHOME |
| 416016530 | MikeMa\_ |  | 875950664 | amyg429 |
| 4510755552 | \_Makada\_ |  | 4256870860 | dina\_ganzert |
| 1067182676 | T\_hale1010 |  | 711407992147730432 | createpositive\_ |
| 282695161 | WayneDupreeShow |  | 742095128664363013 | BuildTheWallFTW |
| 39334221 | pastormarkburns |  | 4826274994 | Lisabridgeman6 |

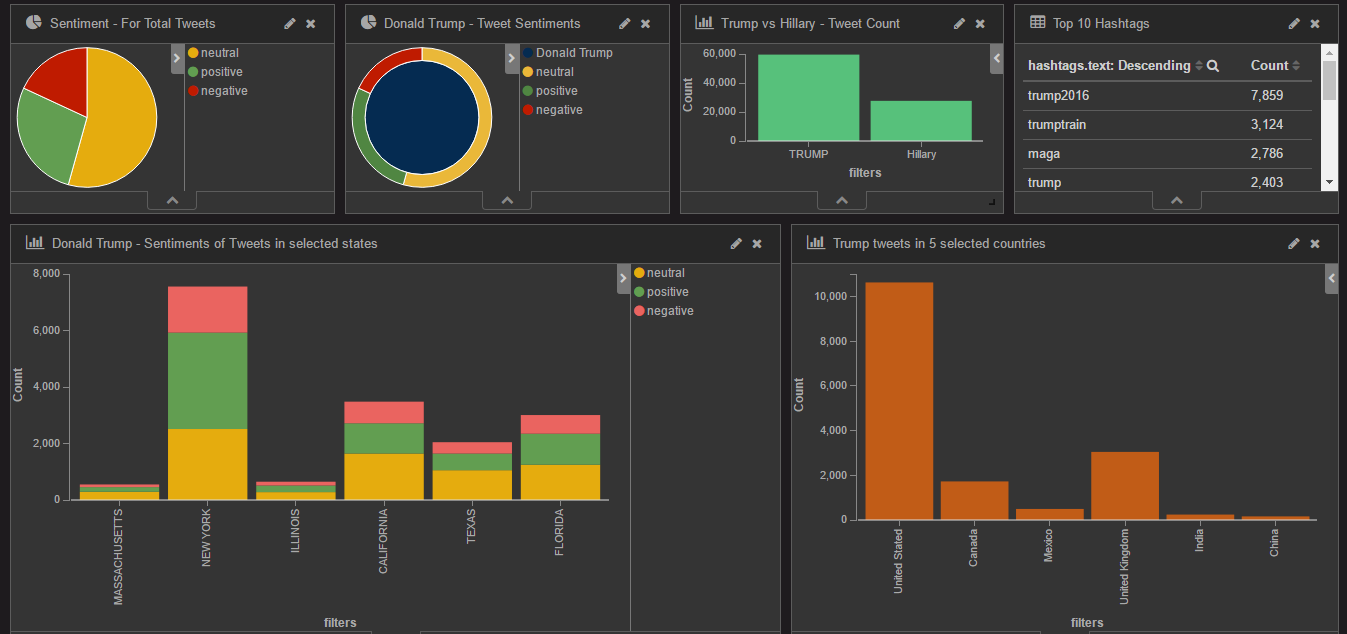
And we also get mutual friends. Table 3.3 is for mutual friends between the followers and friends of USAforTrump2016.

**Table 3.3: Example mutual friends of @USAforTrump2016**

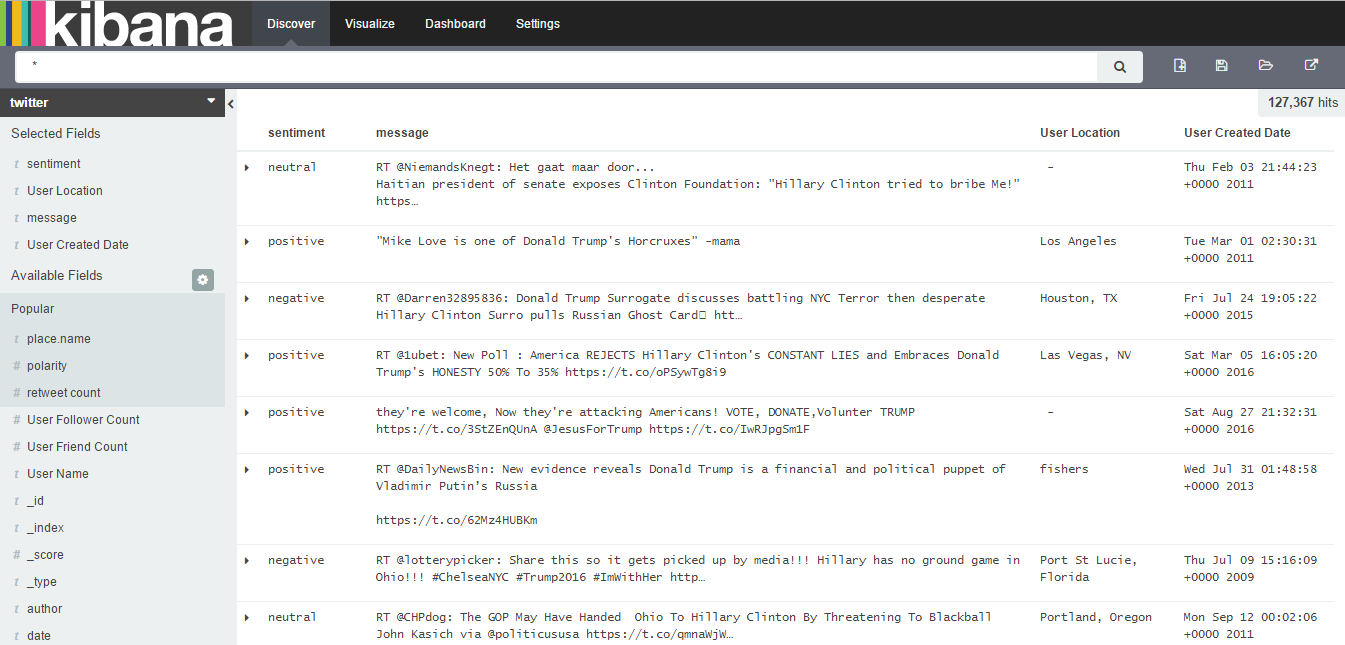
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| ID | SCREEN NAME |  | ID | SCREEN NAME |
| 715595673681338368 | victoriafromDTX |  | 3246686280 | woon04071 |
| 746158854212198402 | trumpflorida16 |  | 1067182676 | T\_hale1010 |
| 711593583644545024 | willspeakout |  | 399745111 | HorseShort |
| 3410972164 | VanessaTreft |  | 299081304 | idebbie1 |
| 166751745 | RyanAFournier |  | 3914047581 | Italians4Trump |
| 735625699550072832 | kirbyREGATED |  | 4625559136 | Danielle9g |
| 743245746389065732 | OHforTrump2016 |  | 2149651044 | I\_AmAmerica |
| 27895304 | heytana |  | 3225760870 | Red\_White\_\_Blue |
| 3656965641 | harrisonterri55 |  | 4430503527 | TrumpPenceUSA1 |
| 3424711689 | TrumpetingTrump |  | 3949709416 | caf514 |
| 3836008983 | PlaysTrumpCard |  | 4437772394 | DirtyBlckPistol |
| 1391035928 | JustinCap\_ |  | 376458353 | RevRiley |
| 2884079135 | drzevetchin |  | 901903482 | OmaTalley1 |
| 91882544 | DineshDSouza |  | 3383049855 | Kotcha301 |
| 2282740273 | Lcourseteach4yo |  | 710990834972233728 | Always\_Trump |
| 3410689588 | FrenchForTrump |  | 729064401823633408 | B4size25paper |
| 3709245494 | USATrustTrump |  | 316401797 | Bohoover |
| 1606436414 | jerrybascuk |  | 710550537444265985 | Billrollins6 |
| 47253057 | RogerJStoneJr |  | 705566834913890305 | joshdunlap1974 |
| 2292074120 | RonSantoFan |  | 4908239480980238 | Joseph1990 |

**Analysis for Business Case. What did you find in the data?**

Further Analysis-Real Time BI Dashboard



From the dashboard we observe that people in new York are talking about Donald the most,while people in Massachusetts and illinios are not interested.We can also see the sentimental anaylsis state wise and the tweet count about each candidate.



Here we can search the tweets according to place,popular user,sentiment, date etc if we want to observe some particular area and concentrate in it.

**Conclusion:**

* People have certain political issue mentioned much more than others, like safety, race.
* Controversial tweets are the key to get more attention within Twitter
* Twitter Data shows that Donald Trump should spend more resources in Massachusetts and Illinois.
* By keeping an eye on the how the trends evolve the campaign can change its tactics.

**Data Limitations:**

* Twitter users could not represent the whole US population
* The typical Twitter user are an 18-29 year-old educated minority with a well-paying job, and is slightly more likely to be male than female.
* Tweets we collected could not represent the opinion for Twitter Users
* Twitter Users who have strong opinion usually contribute more on the tweets.
* The 3 day Time frame limits the information we could collect
* Real time dashboard is our approach to address this issue.