Twitter Sentiment Analysis: U.S. Election 2016

Supervised by: George Kollios, Katherine Zhao

Priya Ayyappan, Sofia Maria Nikolakaki

BOSTON UNIVERSITY

Motivation

The motivation of this project derives from the abundance of U.S. election data in microblogging sites. The continuous political debates due to the upcoming U.S. elections on November 8, 2016 provoke a public reaction that analyzes the performance of political candidates every day. These opinions are reflected in microblogs, where people have easier access to and feel comfortable to express their true opinion. In this project we consider the paradigm of Twitter. Moreover, due to the significance of the U.S. 2016 elections, several public voting polls focus on estimating the voting tendency. The methods used for these predictions are not disclosed, and therefore it is not easy to evaluate their correspondence to reality. This project estimates the results for the most prominent presidential candidates based on the Twitter microblogging system and compares them to publicly available online polls.

Goals

1) Identify the voting intentions of users on Twitter using the expressed sentiment of their tweets.
2) Cluster the users based on their sentiment and geographic location.
3) Identify the voting trends of a state.
4) Compare and visualize the obtained results with online voting polls.

Data Processing

❑ **Data Cleansing ~** Creating a useful dataset

- Removal of noisy data, i.e. advertisements, bad-formatted tweets and repetitions.
- Storing the relevant fields of the .json twitter objects that are required by our implementation on the disk:

  < location, username, account, tweet ID and text >

Before cleansing **~ 5 GB** **(1M tweets)** → **After cleansing** **~ 200 MB** **(60K tweets)**

❑ **Sentiment Analysis ~** *On Tweets (140-limit characters text)*

- Use of the well-known **Bing Liu opinion lexicon,** i.e. dictionary of positive/negative words.
- Perform **Porter Stemming** technique on the tweets and the lexicon
- Find the percentage of positive/negative words per tweet. If strongly positive or strongly negative sentiment exists (80% or more) assign positive or negative sentiment respectively. Otherwise, assign neutral.
- Exploit the revealed powerful sentiment of hashtags, #CruzCrew

❑ **Candidate Extraction ~** *Bernie, BernieSanders,#VoteForBernie*

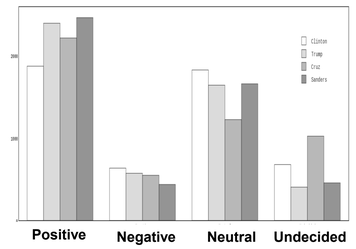
- Create a dictionary with observed versions of a candidate's name.
- A tweet that contains more that one candidates is disregarded.
- A user that has expressed a positive feeling for one or more candidates is considered undecided.

❑ **Location Assignment ~** *Location is not always present*

- Create a dictionary with big cities and the center <lat,lon> of their state
- Correspond tweets that contain information about their location to the <lat,lon> of their respective state.

Results

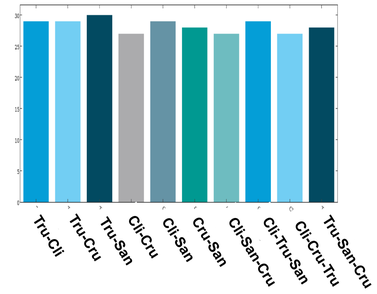
Sentiment Histogram Per Candidate



Predominant candidate for each sentiment:

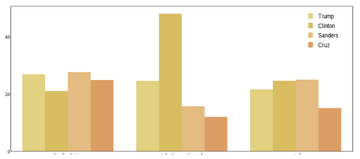
- Sanders – positive (2457 counts)
- Clinton – negative (634 counts)
- Clinton – neutral (1822 counts)
- Cruz – undecided (1021 counts)

People's dilemma



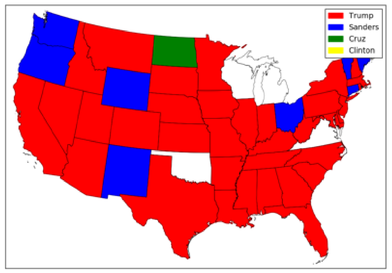
- Most people are undecided between **Trump** and **Sanders** (30 counts)

Comparison between Our Results and Online Voting Polls



- Our results agree with source (3) that **Sanders** has relatively more supporters.
- Our results are more optimistic about **Cruz**.

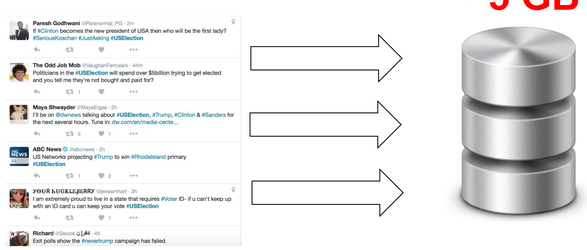
Predominant Candidate per State



- **Trump** it is!

Data Collection

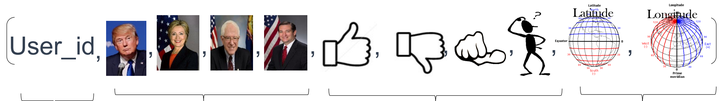
~ 5 GB



- **Twitter Streaming API** setup for 24-hour tweets retrieval
- Queries based on hot election hashtags
  e.g. #HillarySoQualified, #CruzCrew, #FeelTheBern, #VoteTrump
- Tweets arrive in .json format with discrepancies

Methods

- **Feature Vector,** the input to our methods:



unique | 0 or 1: one candidate per feature vector | 0 or 1: one sentiment per feature vector | degrees

- **Classifying** users based on their sentiment for a specific candidate by defining a unique <candidate,sentiment> pair.
- **Clustering** the users based on their sentiment and geographic location <lat,long> using the **K-Means** clustering algorithm. The number of clusters is equal to the number of states presented in our dataset.

Conclusions

- ❑ Twitter generates thousands of political tweets per day, but only 5% of them are useful.
- ❑ Our results are on a very small percentage of the potential supporters of each candidate, and therefore are a rough estimation. Further data collection is necessary for more accuracy.
- ❑ From our analysis we conclude that the political U.S. election race of 2016 will be mainly between Donald Trump and Bernie Sanders.
- ❑ The dominance of Hillary Clinton projected by the voting polls is not observed in our results.

References

1) http://socialmedia-class.org/twittertutorial.html
2) https://twitter.com/
3) https://galeascience.wordpress.com/2016/03/23/us-city-to-state-python-dictionary/