DSCI-644 Development Book

Team Members: Maxime Bost-Brown, Kritin Moondra, Sarah Rote, & Chris Ponge

Created: Spring 2019

Table of Contents

Background:

Social media has become a large part of daily life for Americans and can have an influence on the world around us. In 2016, Twitter may have played a role in the Presidential election between Donald Trump and Hillary Clinton. Using machine learning and data analytics, we will explore the tweets of Donald Trump, Hillary Clinton, and Congress during the time of the election.

Scope:

The scope of this project is to provide analysis of the U.S. 2016 presidential tweets on the levels of a basic twitter user, political analyst, and a data scientist. This includes the tweets of Donald Trump, Hillary Clinton, and Congress members.

Objectives:

- Provide analysis of the U.S. 2016 presidential tweets on 3 levels: basic twitter user, political analyst, and data scientist
- Use a bag-of-words model to classify each tweet as a positive or negative tweet
- Run queries within and across the 3 datasets to obtain information on congressional support, or lack thereof, and draw conclusions based on data analysis
- Access visual and nonvisual statistics to aid in research on political subjects and ideologies
- Allow viewing of the similarities and differences between presidential candidates and congressional members on twitter to help educate the American political atmosphere

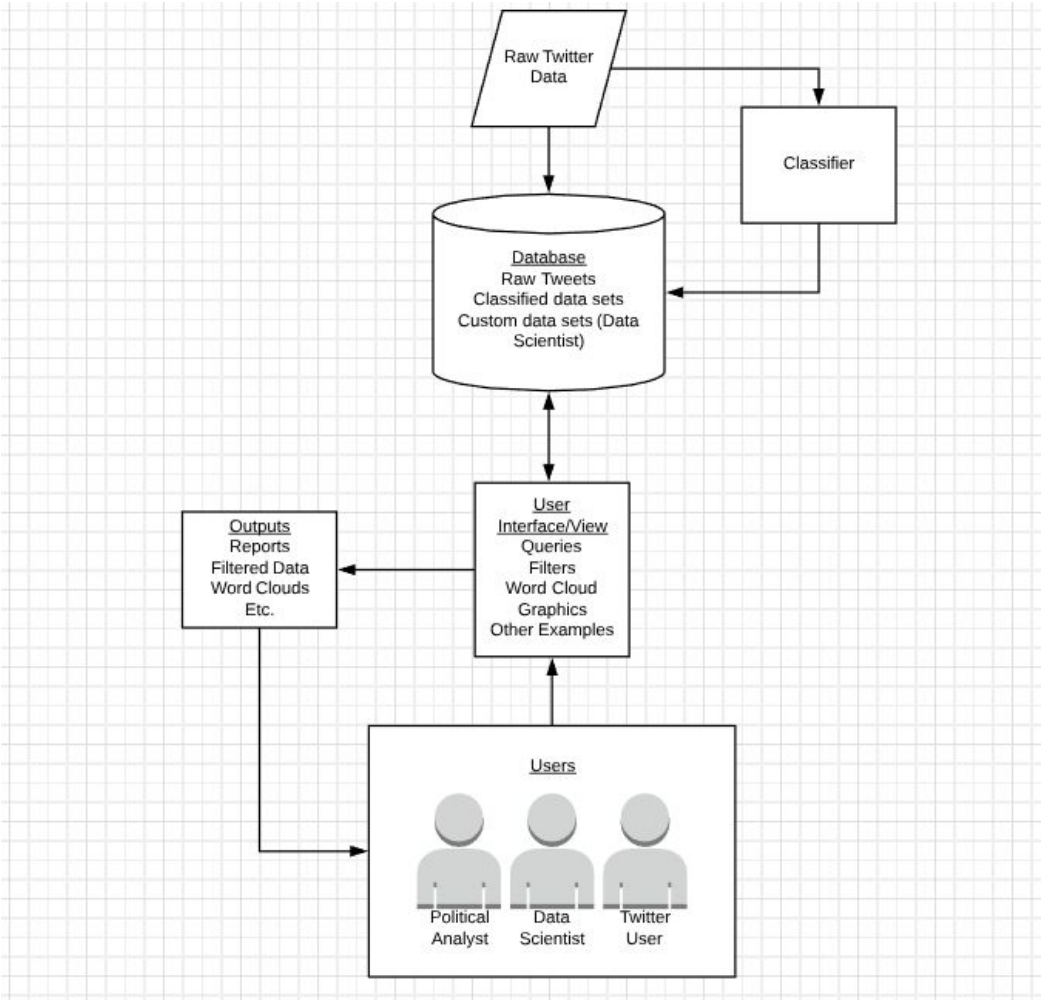
Timeline:

I. **Inception Phase (Weeks 6 & 7)**
   A. Detailed project proposal
   B. Key requirements
II. **Elaboration Phase (Weeks 8 & 9)**
   A. Architecture
III. **Construction Phase (Weeks 10,11, & 12)**
   A. Final Project Delivery
IV. **Transition Phase (Weeks 13 & 14)**
   A. Team Reflection
   B. Presentation

Project Description:

Our team plans to analyze the retweets of Trump and Clinton by Congress to explore if members of Congress retweet their candidate's tweets with positive or negative sentiment. We will also analyze how Congressmen welcome their candidates tweets by analyzing their retweets. This result will help us understand how supportive or critical Congress members are of their respective candidates. Reception of each tweet and their reaction at different intervals of time will also be analyzed. This will help us understand whether the members of Congress are really in sync and touch with their candidates through social media and share their opinion on day to day topics and not on popular topics which have a tendency to become controversial. We will apply machine learning techniques, such as a bag-of-words model to help us do this.

Project Architecture:



Technologies Used:

- Slack: Project Manangement
- Trello: Project Management
- Github: Code Repository
- WhatsApp: Team Communication
- Python
- R

User Stories/Requirements:

1. Data Scientist:
    a. As a data scientist, I want to be able to run queries within and across the three datasets to obtain information on congressional support, or lack thereof, and draw conclusions based on the data analysis.
        i. I want to be able to run queries on the datasets and return objects matching my criteria to be able to do additional analysis on.
            1. Be able to create a subset of tweets based on liberal or conservative sources
            2. Be able to create subsets based on extreme political learnings or moderate. Able to adjust those thresholds at will
            3. Be able to create subsets based on age of congressmen
            4. Be able to create subsets based on gender
        ii. I want to be able to perform a sentiment analysis on the tweet text and classify target tweets as positive/negative/neutral
            1. Be able to send a string of text to a cleaning function that can be forwarded to a classifier
            2. Be able to send cleaned/tokenized text to a classifier and return if the text is positive/negative/neutral.
2. Political Analyst
    a. As a political analyst, I want to be able to access visual and non-visual statistics to aid in research on political subjects and ideologies so that I can interpret and report findings.
        i. I want to know the average number of positive and negative tweets of congressmen above 61(average age of congress) so that i can understand whether they promote positive messages to the public or are cynical towards every event in the society
            1. Get the tweets and the congressmen above the age of 61 from congress.csv file
        ii. I want to be able to look at how often liberal or conservative tweets are retweeted to see how far their message is reaching

1. Be able to check how often a congressman is retweeted
2. Be able to check the total number of retweets/ratio of retweets to total tweets that are put out by congress.
3. Be able to do the same to clinton/trump

iii. Be able to look at the tweets of a person or group of people to see if they tend to spread a positive or negative message

1. Be able to get a total number of overall positive or negative tweets put out by a congressman or trump/clinton
2. Be able to get a proportion of tweets put out by a person that are positive or negative
3. Be able to get a total number of positive/negative tweets posted by a specific group of people.

3. Twitter User
    a. As a twitter user, I want to be able to see the similarities and differences between presidential candidates and congressional members on social media to educate myself on the American political atmosphere.
        i. I want to know the number of positive and negative tweets by Clinton/Trump so that it will help me decide whether she/he is a optimistic or a pessimistic person which will help me in following her/him
            1. The number of both positive and negative tweets should be presented to the user separately,
        ii. I want to be able to look at congress and see if they tend to send positive or negative tweets out
            1. Be able to pick a congressperson or Trump/Clinton or type of congressperson and see how many of their tweets are positive or negative.
        iii. I want to see what congressmen are the most popular by how often they are retweeted
            1. Be able to get the number of retweets that a congressman has.
            2. Be able to get a list of the top congressmen and women by the number of retweets they have.

Layering:

1. Presentation Layer - The front facing part of the application, this is what the user will interact with

2. Business Layer - The meat of the application. This layer will handle all the business rules and send data from the data layer to the front end.

3. Data Layer - The backend of the application. This layer will contain all the database interaction. It will fetch data from the database and send it to the business layer, and take data from the business layer and enter it into the database.

Code Details:

1. Filters
    a. filter_age Gives a dataset only including tweets by people above or below (or exactly) an age
    b. filter_sentiment_type Gives a dataset with only positive, negative, or neutral tweets
    c. filter_sentiment_number Gives a dataset with sentiment valusea abve or below a number (find more extreme sentiment)
    d. filter_gender Gives a dataset with only tweets from specified gender
    e. filter_leaning_tweets Gives a dataset only from liberal or conservative
    f. filter_persons_tweets Gives dataset with only one person's tweets
    g. filter_race_type Gives a dataset containing only the race entered
2. Functions
    a. get_sentiment_count Returns the count of tweets containing that classification of sentiment (+ 0 -)
    b. get_total_tweet_count Returns number of rows in dataset
    c. get_retweets_per_tweet Returns sum of retweet_count divided by number of tweets
    d. get_average_sentiment Returns the average sentiment score in a dataset
    e. get_leaning_score Returns the political leaning score of a person
    f. get_sentiment_frequency returns proportion of tweets that are a particular type of sentiment
    g. get_top_retweeted returns a ranked dictionary of the most retweeted by their retweets_per_tweet and their retweet rates
    h. get_most_conservative returns a dictionary of the most conservative by dw_score and their scores
    i. get_most liberal Same as above but for liberal
    j. get_most_positive returns ranked dictionary of names and average sentiment of most positive average sentiment
    k. get_most_negative same as above for negative

3. Sentiment Analyzer: An analyzer built using Lexicon and word cloud to be able to determine if a tweet has a negative, positive, or neutral connotation.
    a. The program does the following: Puts the three tweet datasets into dataframes. Performs sentiment analysis on the tweets using the nltk.vader tool. This is a lexicon based sentiment analysis trained using social media sources, so we assume it is somewhat applicable. The analysis is added to the dataframes in two forms, the overall score from -1 to 1 showing magnitude of sentiment, as well as an integer score of -1,0,1 (meaning negative positive neutral) showing only direction of sentiment. Called Vader_Score and Trinary_Score.
    b. A shortcoming of this analysis as is is that any new slang terms or created words or hashtags likely won't be interpreted by the classifier so they'll be simply counted as neutral. Therefore, it might miss some of the data.
    c. WE hand classified some data for test of accuracy of classifier. We got about 70% accuracy over fifty random data points from the congress dataset. This seems good enough for our purposes.

Interface:

There are 3 main areas of the interface.

1. Generate Word Clouds
    a. Here you can generate word cloud images of the Clinton data, Trump data, Congress data, Clinton sentiment data, Trump sentiment data, or Congress sentiment data.
2. View Sentiment Data
    a. Here you can view the Clinton sentiment data, Trump sentiment data, or Congress sentiment data. This section allows you to view the raw sentiment data, view just the columns of the data, filter on any or multiple columns, create a histogram of the data based on user parameters, or create a scatterplot of the data based on user parameters
3. View Raw Data
    a. Here you can view the Clinton raw data, Trump raw data, or Congress raw data. This section allows you to view the raw data, view just the columns of the data, filter on any or multiple columns, create a histogram of the data based on user parameters, or create a scatterplot of the data based on user parameters

Deployment:

There are a few steps you must follow in order to download and install the application.

Requirements:

- Python 3.6+

Steps:

1. Go to https://github.com/RIT-DSCI-644/team-project-team-3 and click the green "clone or download" link, then click "download zip".

2. Unzip the file to a location on your local machine.
3. Run the gui2.py file.
    a. Depending on your OS, the path names of the .csv files may need to be changed to include the full path. (This is the case with windows)
    b. You may need to pip install some libraries if you don't already have them:
        i. pip install kivy
        ii. Pip install wxPython
        iii. Pip install wordcloud
        iv. Pip install numpy
        v. Pip install pandas