Social Media Response to Political Ads

Jason Becker, Nalini Chandhi, Brian Schneider

Introduction

Social media and Political Ads are playing more important role in Presidential campaigns than ever before. The type of Ad has a significant impact on how people receive, respond and share to it. The objective of this project is to collect Political Ads data and Social Media response about both Presidential candidates and to investigate any correlation between the Ads and sentiment of social media responses.

Data

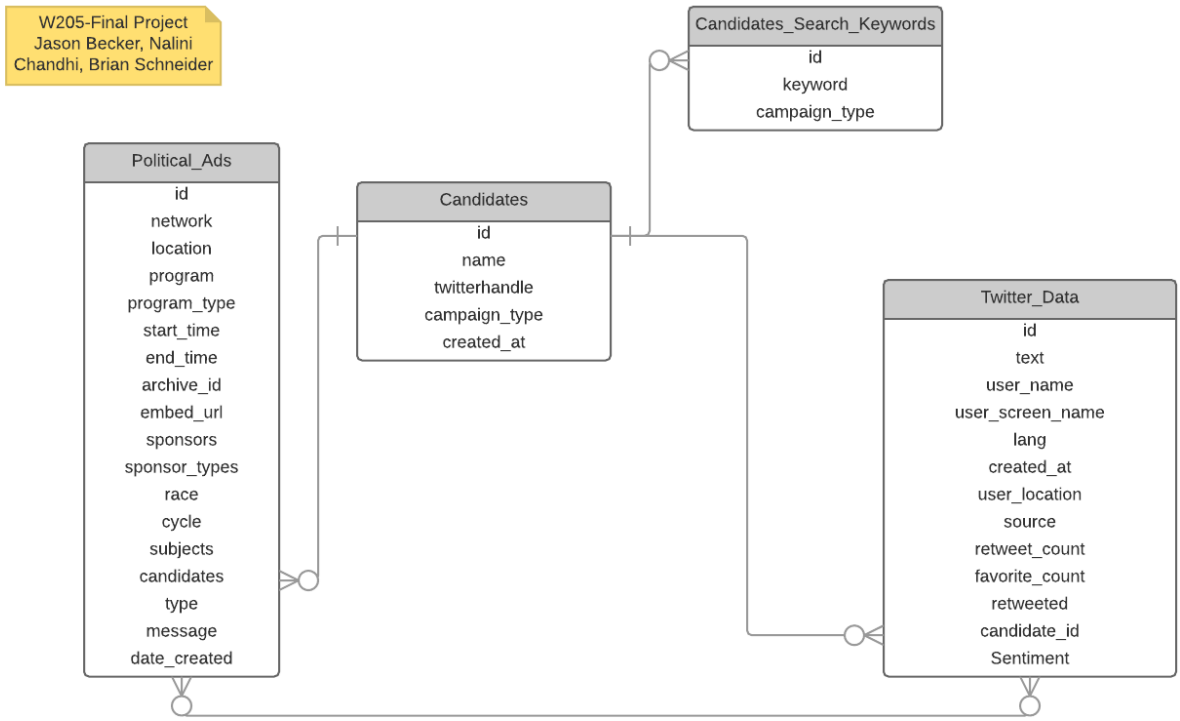
The Political Ads data is available for download at http://politicaladarchive.org/data/ in csv format. We plan to download the complete historic data set initially and load to the data lake. The data is uploaded every month to the website so we plan to build a batch process to upload the same to Data Lake.

We plan to use Twitter search API to filter presidential candidates related key words like 'Hillary Clinton', 'Donald Trump', '@HillaryClinton', '@realDonaldTrump' and collect the tweets into the data lake.

Extensions

We would like to extend this project to include other important events like Presidential debates and positive/negative disclosures from the Candidates or any major information leaks on social media that may change the sentiment (if time permits)

ER Diagram



Technologies

Task	Technologies	Challenges
Data Lake Platform	AWS EC2 instance with EBS volume attached	(Still reviewing other options)
Twitter data collection	Twitter Search API, Tweepy Python library	Twitter API rate limiting (solved by adding proper error handling in the python script).
Ads data collection	Python script	
Sentiment Analysis of Tweets	TextBlob python library	(Still reviewing other options)
Data Joins and queries	Hive/PySpark/SparkSQL	(Still reviewing options)
End user Reports	Tableau	
ER Diagrams	Lucid Charts	
Code Repository, Team collaboration	Github	

Data Challenges

1. Mapping the timeline of Ads and twitter data – Our current plan is to analyze the twitter data after few hours/few days of Ad timeline. Time zone differences in the two data sets make it even more complex but we would like to use location information in both data sets to come up with a better model. We hope to learn and improve the approach as we explore the data.
2. Processing the twitter data to identify fake accounts, Bots, and marketers – Our plan is to begin data analysis with all data and then add as many filters as possible.

Analysis and Reports

1. Sentiment analysis over time (agnostic to advertising) of both candidates on Twitter
    a. How many people are thinking about each candidate?
    b. What is the tone of the tweets relating to each candidate?
        i. Positive
        ii. Negative
    c. How many people tweet the same thing at a specific time frame in response or in addition to a specific candidate?
2. Correlation between sentiment for a Candidate and Political Ads
    a. During a time period for each Political Ad
    b. Based on the location of advertisement and location of people tweeting

Work Plan and Distribution

Task	**Ownership**	**Dates**
Setup the infrastructure for Data Lake		
Write Twitter data load scripts		
Write Political Ad data load scripts		
Write data cleaning & transform scripts		
Write Sentiment analysis scripts		
Build data models		
Build reports		
Monitor data loads & troubleshoot		
Prepare final project presentation		