Political Ads and Social Media Response

Jason Becker, Nalini Chandhi and Brian Schneider

School of Information, University of California, Berkeley
December 7, 2016

jason.becker@ischool.berkeley.edu
nalini.chandhi@ischool.berkeley.edu
brian.schneider@ischool.berkeley.edu

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 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.

Some questions we wanted find answers to:

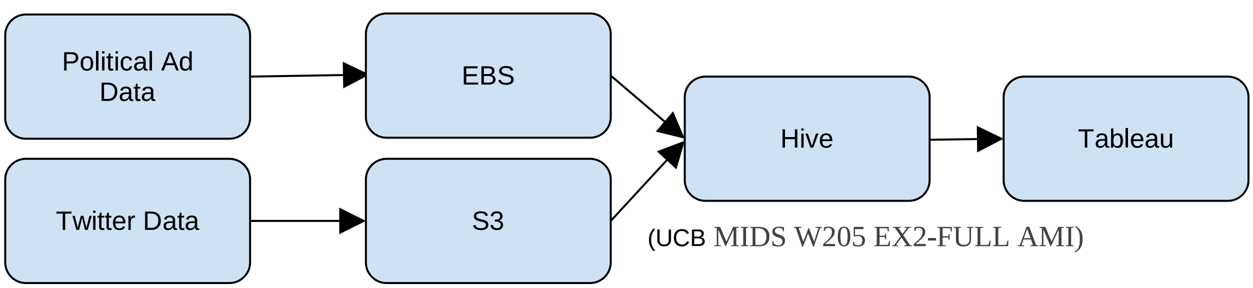
- Does TV advertising meaningfully impact online Twitter sentiment? If so, how much?
- Should political candidates continue to buy ad spots (to motivate Twitter comments)?
- Are some ads more effective at spurring Twitter comments than others?

Data sources

There are two major sources of data

- Political Ad data from http://politicaladarchive.org
- Twitter stream data using Tweepy API

Architecture



Considerations

- We chose S3 because it is scalable and cheap. It is easy to scale as we collect more tweets data
- We chose Hive because of familiarity and easy connectivity to Tableau
- We chose Tableau because of familiarity and ease of report creation. It was a good fit for this use case because of custom geographies feature to draw DMAs.

Scale of the system

- S3 will scale as we stream more data, but it could become expensive. So we can come up with data aging scripts to clean up the old data that may not be needed for analysis
- We can add more EC2 nodes to the cluster to handle bigger data loads. The queries may take longer as the data size grows if we choose not to scale out.

Data Volume

Dataset	Size	Number of Records	Frequency
Political Ad data	246KB	2300 ads per day Total - 114,099	Daily
Twitter data	400 files(60MB each) per day = 24 GB per day 8440 files(60MB each) ~ 400 GB in total	4 million tweets per day Total - 40 million tweets	Stream (but loading to Hive once a day)
DMA shapes data for Tableau	247 MB reduced to 10 MB	4.8 million rows - reduced to 186,000 rows using Alteryx Generalization algorithm	One time process only to be loaded to laptop where Tableau is running
Cities and DMA mapping	1.5 KB	29,000 rows	One time load

Development Objects

The development objects can be divided into three below sections:

- ## Twitter Streaming
  - New application registered at https://apps.twitter.com
  - A Python script using twitter API consumer key and secret and tweepy python library to stream the data and create batches of 10000 tweets in JSON format as files to S3
  - Used ‘Hillary Clinton’ and ‘Donald Trump’ as filter key words to the API
  - Used boto.s3.connection library to connect to S3 from EC2 instance

All Buckets / twitterstorencbsjb

Name	Storage Class	Size	Last Modified
20161020012559.tweets	Standard	36.7 MB	Fri Oct 28 12:01:29 GMT-500 2016
20161020012911.tweets	Standard	36.4 MB	Fri Oct 28 12:01:37 GMT-500 2016
20161020013223.tweets	Standard	36.1 MB	Fri Oct 28 11:52:30 GMT-500 2016
20161020013535.tweets	Standard	41.2 MB	Fri Oct 28 11:55:24 GMT-500 2016

- Ad data
  - A daily cron job (pretest.sh) is created to repeatedly pull in TV ad data via CSV and store it in EBS.
  - A daily cron job (test.sh) is created to update the tables once data was successfully stored.

- Data Load and Transformation
  - A python script is created to load twitter data JSON files from S3 and apply following transformations:
    - Parse JSON to get nested data elements
    - Get location attributes from tweet data and map to DMA (Designated Market Areas)
    - Cast Date Time columns to correct format
    - Get Sentiment of tweet text. Text is split into sentences and scored for both polarity and subjectivity. Then the maximum polarity and subjectivity are retrieved for each tweet.
    - Create the tweet time 5, 10 , 15 minute intervals (used for time attribution during data modeling)

  - A SQL script is created to perform following transformations during Ad Data load:
    - Cast Date Time columns to correct format
    - Refactor Candidate fields to be join-friendly
    - Remove all non-presidential ads
    - Create the Ad time 5, 10 , 15 minute intervals (used for time attribution during data modeling)

  - A Jupyter python notebook using difflib.SequenceMatcher library to match all DMAs from Ad data to DMA master data file to handle small mismatches in text. Some examples of matches are shown below:

Las Vegas, NV	Las Vegas NV
Raleigh-Durham-Fayetteville, NC	Raleigh-Durham (Fayetteville) NC
Tampa-St. Petersburg, FL	Tampa-St. Petersburg (Sarasota) FL
Cleveland, Ohio	Cleveland-Akron (Canton) OH

Cedar Rapids-Waterloo-Iowa City-Dublin, Iowa	Cedar Rapids-Waterloo-Iowa City & Dubuque IA
Philadelphia, PA	Philadelphia PA

- A shell script to get all twitter screen names who have greater than 500 tweets to feed to Bot Scoring python notebook.

- A python notebook to run all the twitter screen names using Butternut python API and get Bot Scores.

- A SQL script to create aggregate tables on Ad Data and Tweets Data for input to Tableau dashboards.

- Important objects and descriptions

Description	**Object Name**
Twitter stream python script	/Twitter_Stream/streamTwitterDataS3.py
Ad data table creation script	/Addata/hive_base_ddl.sql
DMA Match python notebook	/DMA/Match/DMA_Match.ipynb
Twitter data processing python script, Table load shell script, Hive table creation SQL script	/Twitter_Processing/S3_Parser.py /Twitter_Processing/hive_loader.sh /Twitter_Processing/hive_finalTable.sql
Twitter bot user creation shell script, iPython notebook for Bot scoring, Hive table creation SQL script	/TwitterBots/script_to_get_top_users.sh /TwitterBots/TwitterBots_Notebook.ipynb / TwitterBots/twitter_bot_score_table_sql.sql
Hive Analysis table creation SQL script	/Data_Analysis/hive_analysis_tables_queries - v3.sql
Tableau Visualization DMA shape coordinates for polygon drawing	/Tableau_Visualization/DMA_Shapefile.csv

Data Visualization

Tableau is used to connect to Hive using HiveServer. A custom polygon shape file is loaded to Tableau to draw the DMA polygons for visualization. A sample data set for custom polygon is shown below:

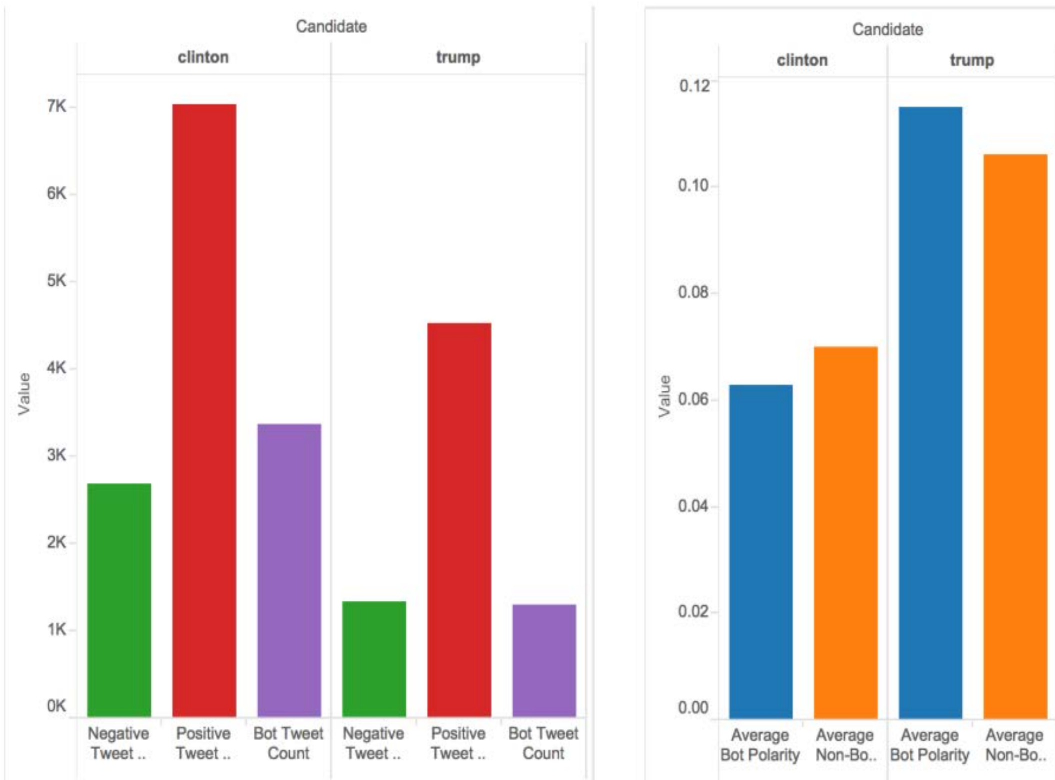
Polygon ID	NAME	ID	SubPolygonID	PointID	Longitude	Latitude
1	Portland-Auburn ME	500	1	1	-70.722602	43.07981
1	Portland-Auburn ME	500	1	2	-70.730029	43.074801
1	Portland-Auburn ME	500	1	3	-70.736067	43.074445
1	Portland-Auburn ME	500	1	4	-70.73855	43.076843
1	Portland-Auburn ME	500	1	5	-70.736944	43.081081
1	Portland-Auburn ME	500	1	6	-70.739073	43.078955
1	Portland-Auburn ME	500	1	7	-70.742816	43.081488
1	Portland-Auburn ME	500	1	8	-70.747211	43.080936

- Political Ads vs Twitter sentiment

  The chart below shows a good correlation between positive ads and positive tweet sentiment and negative ads and negative tweet sentiment:

- Bots by Candidate

The chart below shows that, in the days leading up to the election, Hillary Clinton was far more of a target for negative tweets than Donald Trump. It also shows that bots, in particular, provided negative sentiment towards Clinton, and considerably positive sentiment towards Trump.



- Bots by DMA
  Bot polarity was especially strong in favor of Trump in the Boston, Denver and Phoenix DMAs. Pro-Clinton bot polarity spiked in Washington DC and Philadelphia.

Findings

Based on data analysis shown above, we are able to draw following results:

- Showing ads DOES correlate with Twitter sentiment.
- DMAs do not all respond to ads in the same way.
- We believe that bots did make impact on overall sentiment -- favorably for Trump.

Next Steps

The next steps would be extend this solution to collect other social media data. Currently, the scripts and tables were all created for Hillary Clinton vs Donald Trump election but they can be easily generalized to build a configurable application for any future election. We also expect to build more intelligence and rules into Bot detection and Sentiment Analysis.