Social Media Response to Political Ads Progress Report

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 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.

Infrastructure:

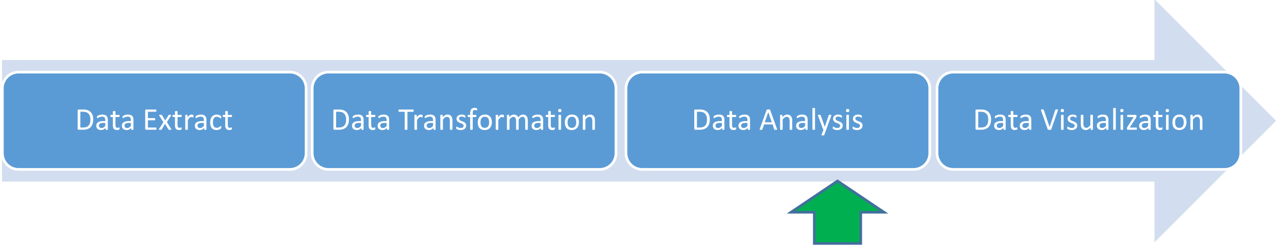
- AWS EC2 instance (UCB AMI) to run the twitter API streaming python script
- S3 to store twitter data files
- Hive on EC2 instance to load and analyze Twitter stream and Political ads data
- Tableau for report development

We started with UCB Spring AMI but had issues upgrading Python to version 2.7 to get Pandas working. So we moved to the new Ex2 AMI and have all components working as needed.

Development plan:

- **Data Extract**
  - Stream and Store Twitter data for both presidential candidates in JSON format in S3 files (10000 records per file).
  - Download Political Ad data in csv format
- **Data Transformation**
  - Parse Twitter JSON data using Pandas to fetch columns of interest and load to Hive table
  - Run sentiment analysis for each tweet text and add Sentiment scores as columns
  - Transform date and location columns to match in both data sets
- **Data Analysis**

    - Join Twitter data table with Political Ad data table
    - Apply necessary filters to clean the data
- **Data Visualization**
    - Create Tableau dashboards to visualize the results



Next Steps:

- Complete building data analysis scripts
- Create Tableau dashboards for visualization
- Bring all moving pieces together, clean up scripts and save to GitHub repository
- Prepare final project submission document and the class presentation
- Try other sentiment analysis libraries if time permits
- Try to identify non-human twitter users and report the results

Challenges:

- We initially filled our 100 GB instance with twitter stream data in less than one week so quickly moved to S3 for data storage
- The date format in two data sets is completely different
- Location details in twitter data are not complete