Social Sensing: Project Proposal

Kurt Davis
Computer Science
University of Notre Dame

George Krug
Computer Science
University of Notre Dame

Eoghan Martin
ESTEEM
University of Notre Dame

***Abstract*—Using political hashtags, the US political landscape will be analyzed by geolocation and correlated with large media events throughout the course of the US presidential campaign over the last year. Data sets from various cross sections of the presidential campaign will be analysed to predict hypothetical election outcomes at those specific times. This model will then be applied to current Twitter data to measure relative political outlook.**

Twitter will be the main source of data and semantics will be deciphered so as to quantify how democrat-oriented or republican-oriented a tweet is directed. Similar to PageRank, a value will be assigned to the likelihood of a tweet being for Donald Trump or Hillary Clinton. Using semantic analysis, tweets that are against each candidate will also deliver insightful results.

I. Overview of Project

This system will be split into 3 main sections:
1. Get and clean data.
2. Semantic coefficient computation.
3. Final result determination.

1. The data for this system will be acquired from 2 main sources. Firstly, from Notre Dames Apollo data collections. This will consist of presidential debates data. The second, source will be from current data. To acquire this data a Twitter crawler must be developed. This crawler will gather current data relative to political opinion. To do this keywords and geo-locations will be used.

2. Stage 2 of the system is the most complex. This stage involves computing a semantic coefficient to determine how Democratic or Republican a tweet is. This token analysis will be developed in a layered fashion. Firstly, keywords will be analysed to calculate the coefficient. Then punctuation and grammar will be incorporated. Finally, semantic analysis must be incorporated.

3. The final result determination involves accumulating the semantic coefficients and splitting tweets into geo-locations.

The output of this system will be a quantification of the political leanings of a specified, surveyed area that can be used in comparisons both spatially and temporally.

The Tweepy library will be used with Python to build the Twitter crawler. The underlying computation for this project will also be developed using Python. The results will be displayed using Pandas as a visualisation tool.

Using the result from the program pipeline, it will be possible to effectively measure the political leanings or sentiment of a given area. The area is specified while collecting the data sets by filtering either live data or JSON data files using location data, or if possible using probabilistic analysis of location based on content.

II. Data Sets

A. Prospective Counties for Survey

MI- Calhoun, Eaton, Livingston, Kent
WI- Winnebago, Sauk, Columbia
PA- Erie, Northampton, Allegheny

B. Prospective Data Sets

Historic: Apollo data sets on the political debates of 2016.
Current: Current Twitter data discussing/relating to the political climate in the U.S.

III. State of the Art

A. A System for Real-Time Sentiment Analysis of 2012 U.S. Presidential Election Cycle

According to the article, the researchers attempted to use real time sentiment analysis on live Twitter stream data to gain insight into the political trends leading up to the US presidential election in 2012. Their data processing and sentiment model aimed to detect the constantly evolving national political mode by responding to real time events and breaking news.

B. Subgroup Detector: A System for Detecting Subgroups in Online Discussions

Authors: Amjad Abu-Jbara, Dragomir Radev
Offering a unique approach to analyzing sentiment, the researchers decided to collect data from online political forums and then predicted user opinions on the subject matter. In addition they were able to successfully group users based on their specific opinions related to the topic of discussion.

C. Analyzing Twitter Sentiment of the 2016 Presidential Candidates

Delenn Chin, Anna Zappone, Jessica Zhao
https://web.stanford.edu/ jesszhao/files/twitterSentiment.pdf
The keywords used for this study were politics, political candidates, or the full name of a 2016 presidential candidates. An interesting approach taken here involves the mapping of

emojis for sentiment analysis.

http://www.semantic-visions.com/

Semantic-visions is a very fascinating startup that uses geo-political and sentiment analysis to provide real time risk analysis for companies in a variety of industries.

D. "What Data Analysis Tells Us About the U.S. Presidential Election"

https://www.technologyreview.com/s/602742/what-data-analysis-tells-us-about-the-us-presidential-election/

Author: Manju Bansal

Instead of using Twitter to gauge political sentiment, the researchers used analyzed content published by political news media sites. The reasoning behind this approach was to achieve better context by using entire news articles versus individual tweets. The sentiment analysis used a formal structure to detect positive, negative, and neutral opinions based on the intensity and frequency of various words.