Social Sensing: Midterm Report

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 can be analyzed by geo-location and user influence and compared with current Twitter sentiment about the President of the United States. Data sets from various cross sections of the presidential campaign will be analyzed to understand electorate outlook at those specific times. This model will then be applied to current Twitter data to measure relative political outlook today.**

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. 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.

Using these aggregated results, geo-locations can be filtered, giving further insight. The influence of Twitter users per location that are for each candidate will then be weighed, giving a further level of insight.

Using the geo-location filters and influence ratings from the data for each side of the campaign, a comparison can be extracted showing how much public opinion has changed or otherwise towards the current administration and election candidates.

I. Overview of Project

The system is split into two fundamental steps. Firstly, the sentiment analysis and aggregation step as seen in Fig. 1, and secondly, the processes that involve influence analysis and geo-location analysis from the previously aggregated tweets.

A. Step 1

This system will be split into 3 main sections and is described in Fig. 1:

1) Get and clean data.
2) Semantic coefficient computation.
3) Final aggregation and output.

1) The data for this system is acquired from 2 main sources. Firstly, from the University of Notre Dames Apollo data collections. This consists of presidential debate data as well as data from election day and the presidential inauguration. The second source comes from current data. To acquire this data a Twitter crawler has been developed. This crawler gathers current data relative to political opinion. To do this, keywords are 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 analyzed to calculate the coefficient. Then punctuation and grammar will be incorporated. Finally, semantic analysis must be incorporated. The stage is described by the preprocessing, match tweet to candidate, and sentiment model sections shown in Fig. 1.
3) The final result aggregation involves accumulating the semantic coefficients and splitting tweets into Republican and Democratic tweets.

B. Step 2

Step 2 is made up of 2 stages; geo analysis and influence analysis. Ultimately, by geographically sorting the data, the results will yield higher resolution insights. As well as that, with the ever growing problem of bots on the internet, and especially bots on social media, it is hard to determine if a pure count of for and against tweets is a fair assessment of Twitter's views. To combat this, an influence parameter is computed for each Twitter user.

1) Geo Analysis: The sentiment analysis and tweet aggregation is split with respect to geolocation. This helps gain a better insight into specific geolocations hence providing for a higher resolution oversight of the US public opinion towards the presidential candidates at these times.

To improve the readings from live data even further, analysis of multiple samples from the same location allows for spatial-temporal tracking of political trends. Applications leveraging spatial-temporal analysis could verify claims about political climate or provide more precise political projections.

One difficulty experienced with geolocated tweet collection is the relative sparsity of tweets that include location data. To assist in more rapid data collection, later iterations of the web crawler component could potentially leverage an assumption to simplify tweet location: A twitter user, while not typically disclosing tweet location, may be more likely to list their area of residence, and by extension their voting location. Since analysis of tweets will attempt to score voting influence, a Twitter user's voting location is the more relevant location metric. Thus, collecting tweets by keyword and filtering by user location may result in faster data set collection.

2) Influence Analysis: Sentiment analysis must also account for relative influence of each tweet. Some Twitter users have a higher influence than others. As a result, it is possible that even if one party delegate has less tweets in favor of them, they may command a wider reach from their Twitter supporters. As an example, the Twitter accounts of large media corporations command a huge amount of influence on public opinion. The influence of a user with under 100

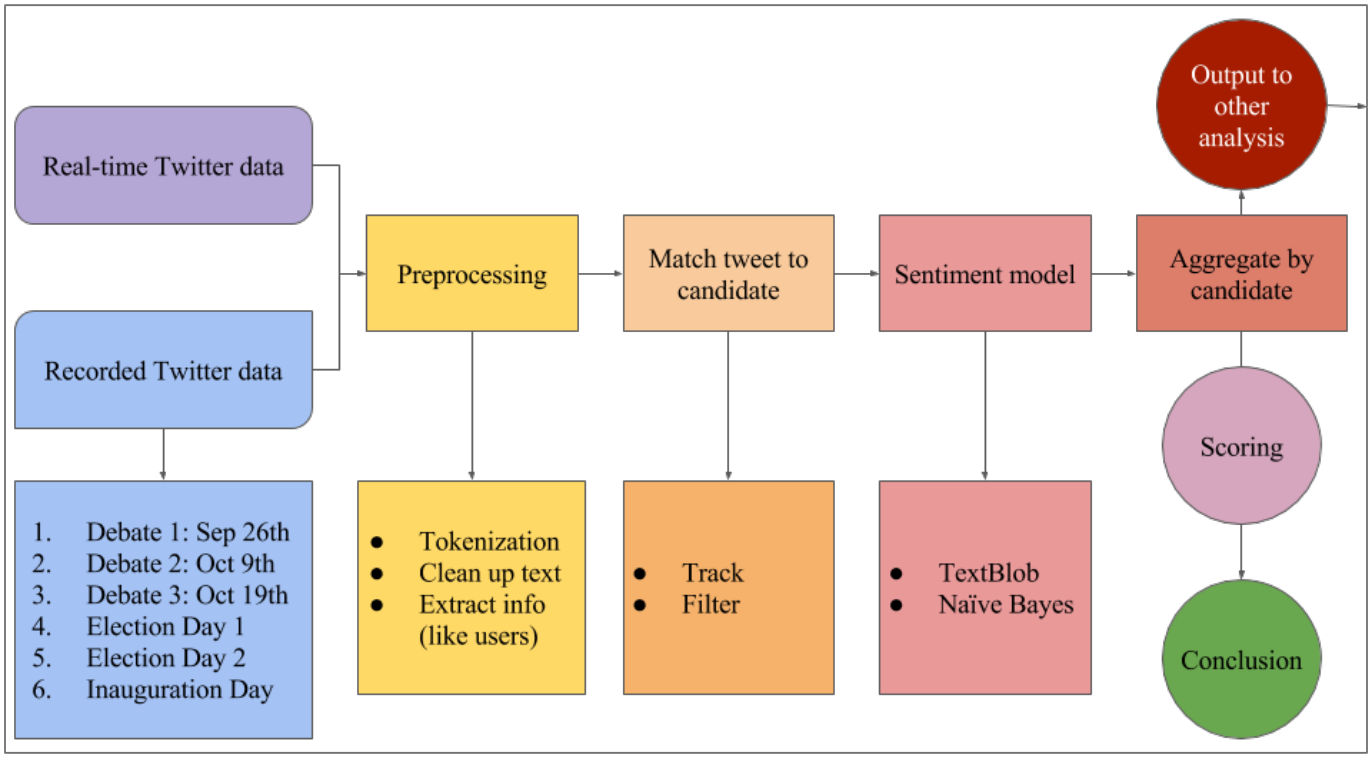

Fig. 1. Step 1: Sentiment analysis and aggregation of Tweets.

followers, however, commands significantly less impact on the candidate's public support. This phenomenon is displayed in Fig. 2. Camp A and Camp B can represent the Democratic and Republican parties. It is possible that camp A has a larger amount of positive tweets about them but as a result of the influence from camp B tweeters, the geographic region highlighted in green is considered to be a camp B region.

An influence rating can be built for each side using:

- Page rank of users: Similar to Google's PageRank, a rank can be assigned to the number of edges a user has within the Twitter social network. This depends on the amount of followers a user has, as well as retweets, interactions, following, and even liked tweets.
- Betweenness centrality: This is a difficult one to measure and may tie back into a user's PageRank. By measuring how short a path is from a user to many users, their betweenness centrality can be determined.
- Ratio of retweets to mentions (talked about value): If a user is retweeted many times, it does not mean that they are as popular as a person who is mentioned a lot of times. Each interaction on Twitter carries a varied amount of influential value. By determining these values and associating them with users, a score can be generated for their influence amongst their network.

C. Output and Technologies

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 visualization 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 using probabilistic analysis of location based on content.

II. Data Sets

A. Counties for Survey

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

B. Data Sets

1) Historic: Apollo data sets on the political debates of 2016.
   - Debate 1: Sep 26th
   - Debate 2: Oct 9th
   - Debate 3: Oct 19th
   - Election Day 1
   - Election Day 2
   - Inauguration Day

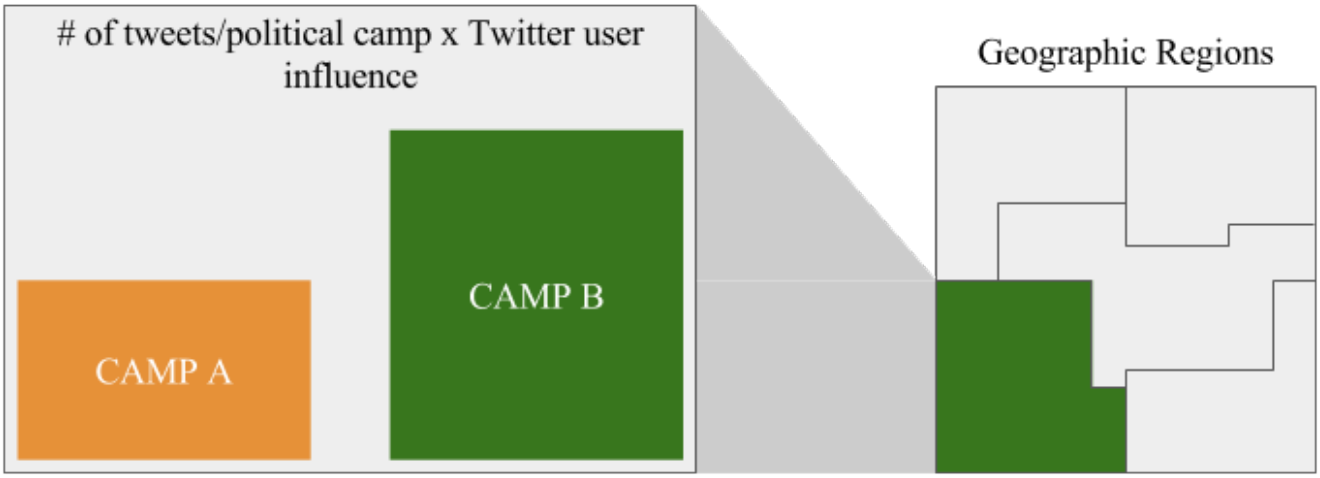

Fig. 2. User influence affects the sentiment analysis model per geography.

2) 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

Authors: Delenn Chin, Anna Zappone, Jessica Zhao

URL: 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"

URL: 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.

IV. Problems

Political tweets in particular, tend to be quite sarcastic. This makes semantic analysis of these tweets very difficult.

It is very common for tweets to not contain geolocation data. As a result, geolocations will have to be inferred in some cases from either the tweet content or the Twitter user's profile location.

Twitter users are only one cross section of populous. The US electorate is made up of many demographics outside of Twitter users. It is important to keep this in mind when analyzing the results.

Twitter only exposes data on a 1 week sliding window. As a result, it is not possible to get real time information about user profiles. This may be an issue when calculating the influence of a user. Their profile will have to be crawled and the data collected will be from a different time to the time that they tweeted about the election.

The size of the data is very large, in some cases over 1GB. As a result, it is necessary to divide and conquer.

V. Milestone 1

As can be seen Fig. 4, the first milestone involves creating a Twitter crawler and collecting real-time tweets relating to the

President of the United States. This crawler was built using Tweepy and collects data with appropriate tags. The crawler was used to collect a sample data set of 500 tweets from across the US relating to the President of the United States. This data is then used in calculating our first early results for semantic analysis and comparison.

The Apollo data sets were also acquired and a GitHub repository was created with a branch for the Twitter crawler and a branch for the semantic analysis model.

VI. Milestone 2

The next milestone involved building a basic semantic analysis model through the use of TextBlob. TextBlob uses a Naive Bayes algorithm to determine whether a statement is positive or negative, returning a 1 for positive and -1 for negative. Using this model and a selection of tweets from both the campaign data sets and realtime data collected, some early results were determined and compared as shown in Fig. 3.

In calculating these results, it was found that large data sets caused a significant slow down in the computation time. MPI for Python was investigated to divide and conquer the computation. Using MPI on the CSE cluster means that the computation can be split across multiple machines, putting less stress on one machine's RAM availability. This is however a difficult process and involves a lot of unstructured local storage of data. The next solution investigated was the use of Tableau or a structured database. Tableau allows for the speedy access and display of data and is currently still being investigated. Future plans involve using Tableau in conjunction with MPI for Python. This means that data can be accessed and stored in a structured way and computation can still be split across multiple machines for speed. Tableau also provides the added benefit of great visualization tools.

A. Results

```
C:\Users\
There are 141 tweets for Donald Trump.
There are 121 tweets for Hillary Clinton.
Score for Hillary: 0
Score for Donald: 20

C:\Users\
There are 120 tweets for Donald Trump.
There are 43 tweets for Hillary Clinton.
Score for Hillary: 0
Score for Donald: 77
```

Fig. 3. First results for comparison using a selection of tweets from throughout the campaign vs. a section of tweets from the week of 20th March 2017.

The current model covers the all stages of step 1 of this project as seen in Fig. 1. The data is acquired, the semantics of that data are calculated, and the results are aggregated into a Donald Trump and Hillary Clinton data structure. These data structures are currently printed to a file.

A naive scoring mechanism was also implemented in the results shown in Fig. 3. This scoring mechanism determines how popular a candidate is while taking into account how negative their supporters are. If a tweet is for a candidate, 1 is added to their score. However, if a tweet is negative towards a candidate, 1 is removed from the candidate receiving the negativity of the tweet and 1 is added to the other candidate's score. This will be further developed to determine which candidate has the most supporters that tweet negatively about the opposition.

The semantic model used here will have to be improved for future iterations. As well as this, the aggregated tweets will be added to a data base once they have been determined to be for a certain candidate. The data in this data base can then be used to split the results further into geographic groups and generate influence scores for users.

The project and results can be viewed at https://github.com/eoghanmartin/SocialSensingProject

VII. References

[1] Title Bots and Automation over Twitter during the First U.S. Presidential Debate - Philip N. Howard, Oxford University, URL: http://politicalbots.org/?p=711

[2] Title: Watchdog to launch inquiry into misuse of data in politics, Author: The Guardian, 4th March 2017, URL: theguardian.com/technology/2017/mar/04/cambridge-analytics-data-brexit-trump

[3] Title: Election 2016ŁŁDebate Three on Twitter, Author: John Swain, URL: medium.com/@swainjo/election-2016-debate-three-on-twitter-4fc5723a3872#.84wt44ak5

[4] Title: Twitter Conversation Performance Measures, Author: John Swain, URL: medium.com/@swainjo/twitter-conversation-performance-measures-c51cf718e18f#.uhqgpgu79

[5] Title: Analyzing Twitter Sentiment of the 2016 Presidential Candidates, Author: Delenn Chin, Anna Zappone, and Jessica Zhao, URL: web.stanford.edu/ jesszhao/files/twitterSentiment.pdf

[6] Title: Trump's Twitter debate lead was 'swelled by bots', Author: Shiroma Silva, URL: bbc.com/news/technology-37684418

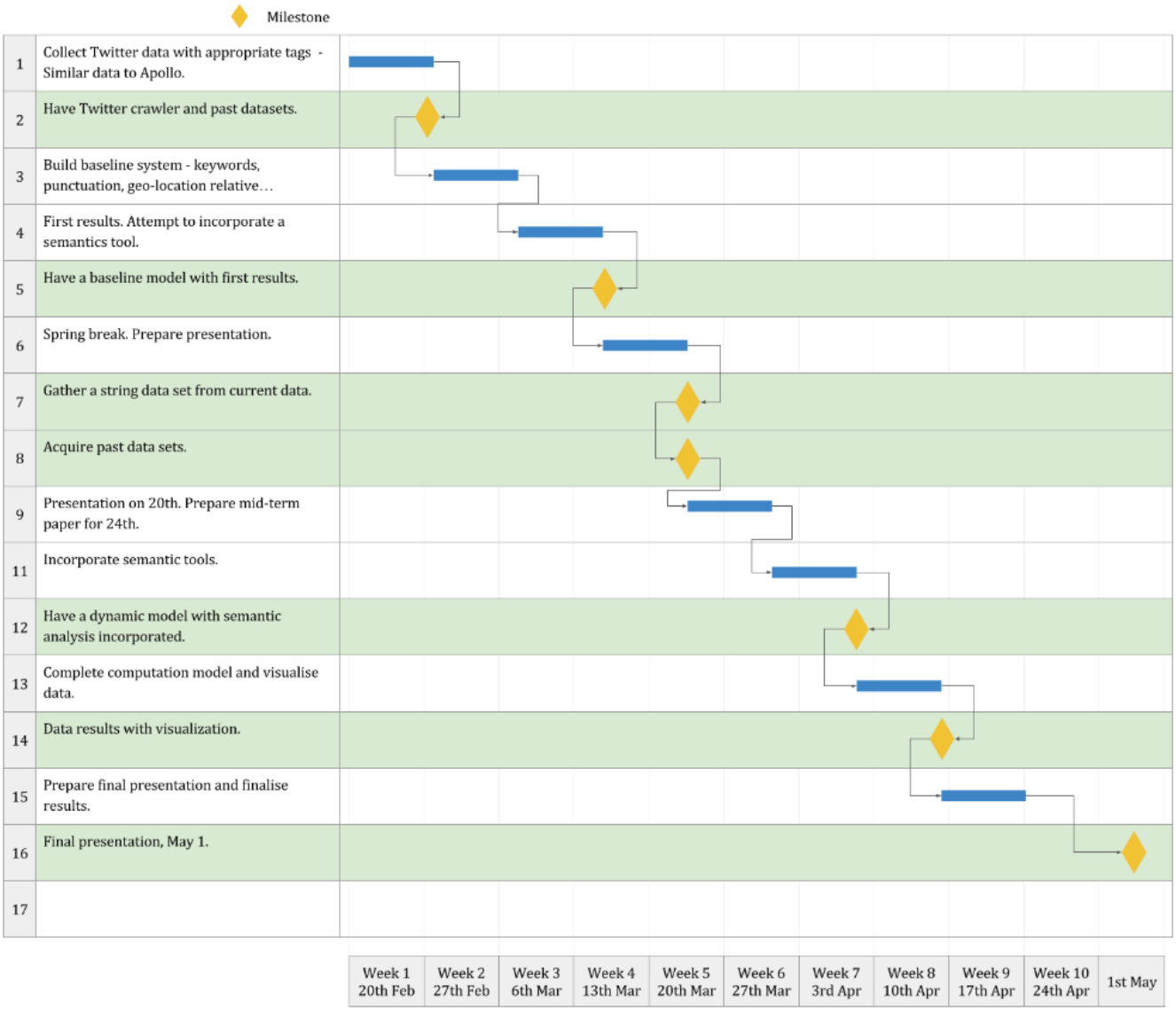

Fig. 4. Milestones Gantt chart.