Presidential Campaign Analysis 2016

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 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 are 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 sentiment will be deciphered so as to quantify how democrat-oriented or republican-oriented a tweet is. A value will be assigned to the likelihood of a tweet being for Donald Trump or Hillary Clinton. Using sentiment 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 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. Problem Statement

Predicting the outcome of the 2016 Presidential Election was mostly inaccurate and consideration for dormant or silence Twitter users was not considered. By taking influence of a Twitter account into consideration, a more accurate prediction model can be built.

II. 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.

III. Solution

The system is split into 6 main steps as shown in Fig. 1.

A. Step 1

The data for this system is acquired from 2 main sources. Firstly, from the University of Notre Dames Apollo data collections, in particular it uses the presidential debate data. The second source comes from current data, acquired by a Twitter crawler has been developed. This crawler gathers current data related to political opinion. The method used to collect tweets eschews the use of geolocation data, as most users will not include such data, instead looking to profile location, which is always present in tweets received by the crawler.

As a result of the unique location filtering, the Tweepy API is used to filter by search keyword (i.e. Trump or POTUS), and tweets sent from the Twitter pipeline are then filtered by comparing listed user locations with a set of strings that could possibly describe a specific place/city. For example, when searching for New York City, possible search strings might

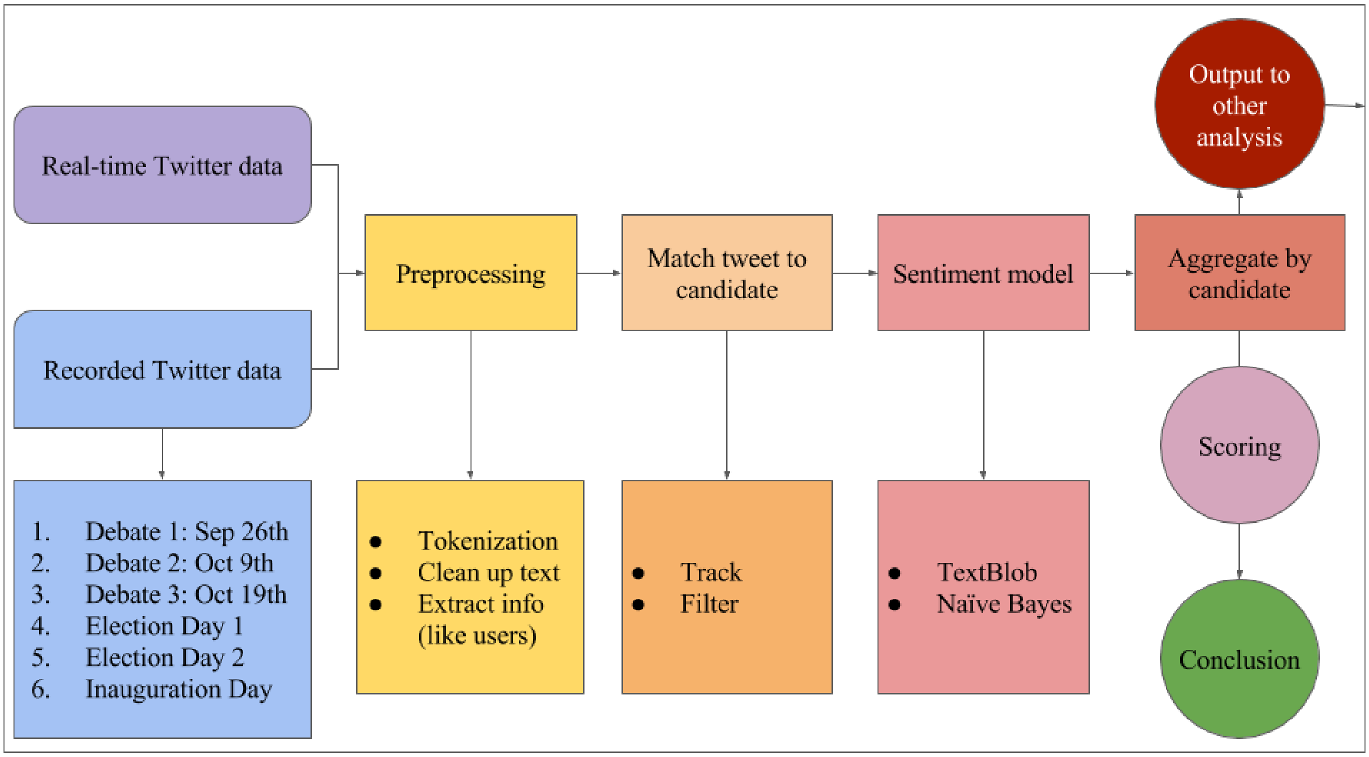


Fig. 1. Sentiment analysis and aggregation of Tweets pipeline.

be New York City or New York, NY. Twitters site offers autocomplete when filling in profile information, providing likely possibilities for the crawler to search with (The latter of the above search strings is one such autocompleted possibility). However, search strings must be particular enough as to not gather tweets from other similarly named locations. For the same example as above, New York might not produce accurate results, as it could pull in tweets from the entire state of New York instead of just the city. The result of this filtering method affords less sparsity to data relative to the geolocation method, but is still not a perfect solution. Whereas geolocation data is unlikely to be spoofed, some users might list their profile locations inaccurately. However, the increased data gained from this method was decided to outweigh the slight potential for greater inaccuracy.

B. Step 2 and 3

The tweet text is converted to lowercase and any cleaned up. To classify which candidate the tweets are about, a range of keywords and hashtags are used. The keywords are shown in Tab. I.

C. Step 4

This stage involves computing a sentiment polarity to determine how Democratic or Republican a tweet is. This stage also involves acquiring a influence value for each Twitter user. To

TABLE I
CANDIDATE CLASSIFICATION KEYWORDS.

Trump	Clinton
#donaldtrump2016	#notmypresident
#trump2016	#democrat
#makeamericasafeagain	#imwithher
#trump2016	#nevertrump
#trumptrain	#feminism
#makeamericagreatagain	#studentloandebt
#draintheswamp	#studentloanforgiveness
#hillaryemails	#climatechange
#neverhillary	#globalwarming
#republicans	#istandwithpp
#hillaryforprison2016	#BlackLivesMatter
#crookedhillary	#campaignzero
#hillaryforprison	#stopgunviolence
#alllivesmatter	

calculate the sentiment of each tweet, the sentiment.polarity function is used from the TextBlob library.

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 followers, however, commands significantly less impact on the candidate's public support. Understanding who dormant or silent Twitter users choose to follow on Twitter can give an indication of their political preference. If there is a large amount of silent Twitter users following influential users on one side of the campaign, that may suggest that there is a bigger following for that side, even without actively tweeting their views. 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 by calculating the ratio between the users indegrees and outdegrees. A Twitter users indegrees equates to the number of followers that they have. Their outdegrees is the number of people that they follow. A user is considered to be influencial when their indegrees outweighs their outdegrees. This is the assumption used in calculating our influential value:

```
follower_following_ratio = self.followers/
self.following
```

D. Step 5

The calculated sentiment and influence values are then appended to the tweet JSON files and are exported for visualization. While doing that, the values are aggregated and some calculations can be computed. Tweets with negative polarities most often reflect negative opinions about a particular candidate. Therefore, whenever Donald Trump is mentioned in a tweet with a negative polarity, the sentiment score for Hillary Clinton will be incremented, and vice versa whenever there is there is a negative opinion concerning Clinton.

Also, in order to account for the frequency bias that favors Trump who is the subject of many more tweets, only strong positive polarities will be counted as positive opinions in order to help normalize the fact that Clintons average polarity is greater than that of Trump. Therefore, whenever a candidate is mentioned in a tweet with a polarity greater than 0.3, his or her sentiment score will be incremented by one. Ultimately, each candidates sentiment score is divided by the total number of opinionated political tweets to calculate the final favorability percentages.

E. Step 6

The final step involves importing and visualizing the data.

IV. Output and Technologies

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

The Tweepy library is used with Python to build the Twitter crawler. The underlying computation for this project is also developed using Python. The results are displayed using Tableau as a visualization tool.

The system pipeline is shown in Fig. 3. Documentation for this system can be seen at github.com/eoghanmartin/SocialSensingProject. The data is either acquired from Apollo or using our Twitter crawler. Large files are then indexed so as to make computation easier on smaller machines. From there, user information is gathered and an influence for each user is recorded. These influence values are used in the sentiment analysis and the sentiment and influence values for each tweet are embedded in that tweets JSON elements. The outputted JSON is imported to Tableau and visualizations are generated to display insights.

A. Data Sets

The sentiment analysis and tweet aggregation is split with respect to geolocation. This helps gain a better insight into specific locations 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. 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 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 results in faster data set collection.

Prospective locations for survey:

- CA - San Francisco
- IL - Chicago
- NY - New York City
- PA - Erie
- OK - Oklahoma City

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.

V. Results and Evaluation

According to the RCP average 42.8% favorable and 53.5% unfavorable for the current administration.

Tweets with weak polarities that reflect ambiguous opinions or unbiased reporting are discarded. This accounts for

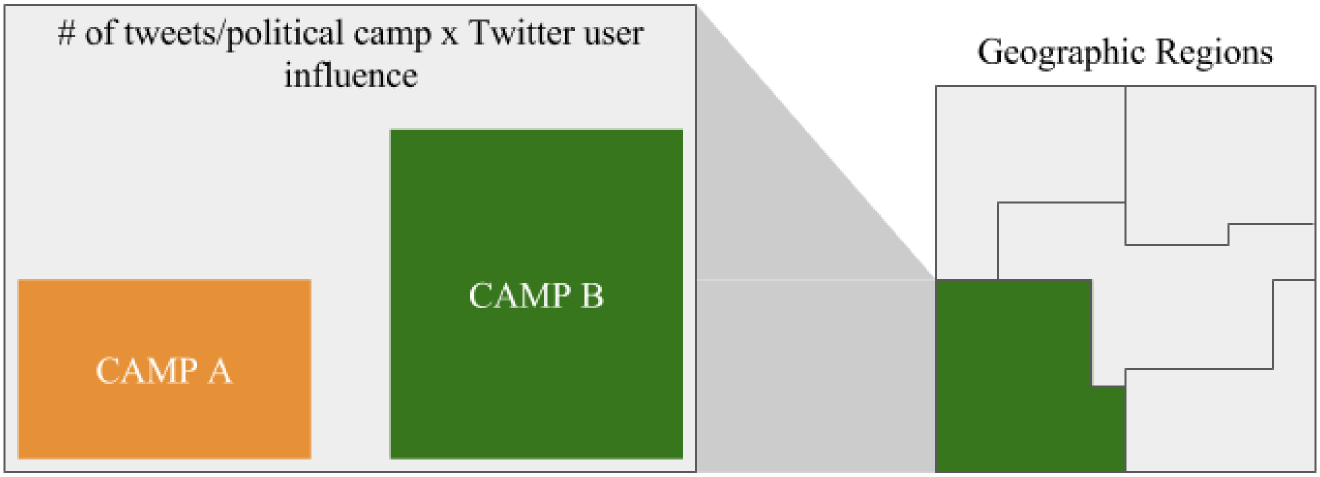


Fig. 2. User influence affects the sentiment analysis model per geography.

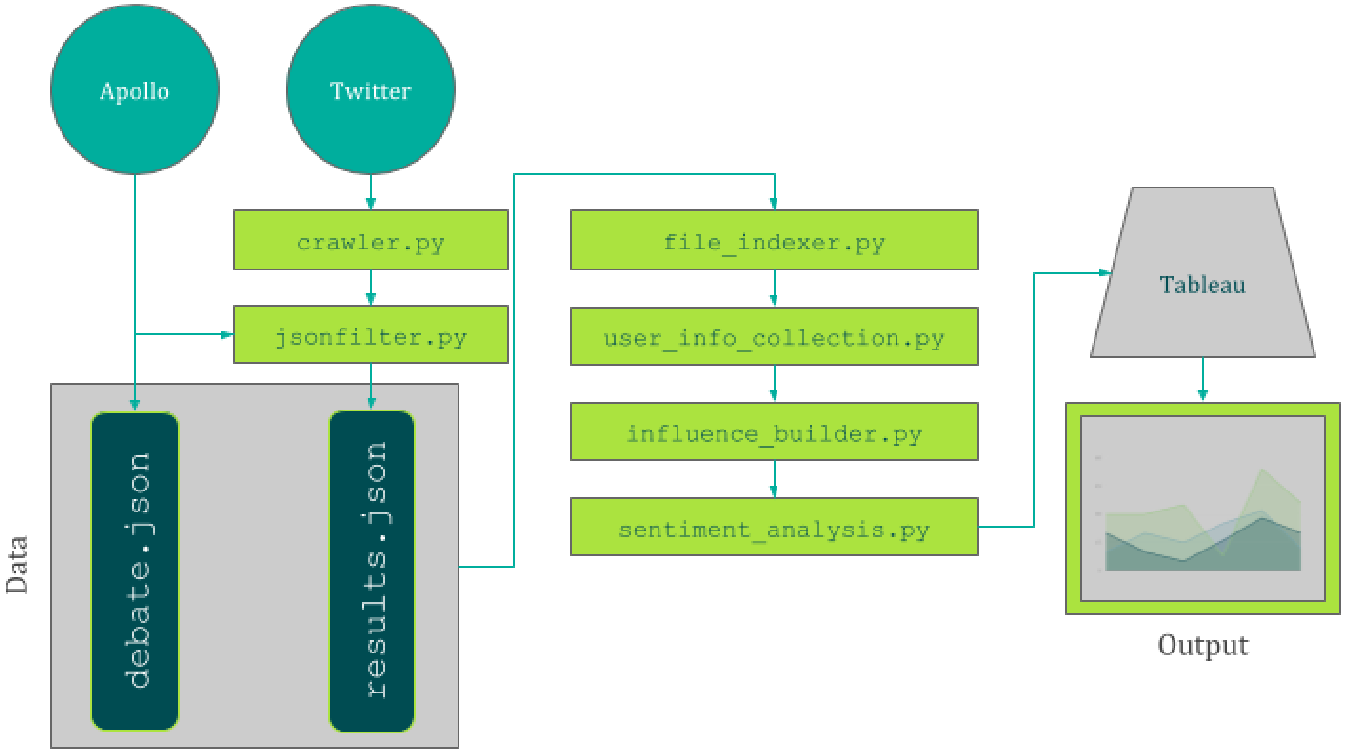
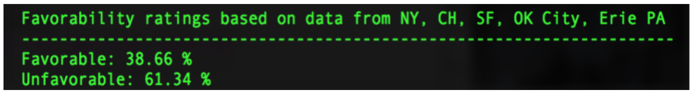


Fig. 3. System pipeline.



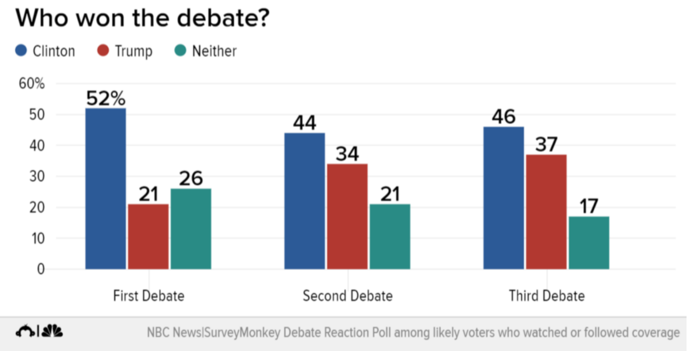
```
Favorability ratings based on data from NY, CH, SF, OK City, Erie PA
---------------------------------------------------------------------
Favorable: 38.66 %
Unfavorable: 61.34 %
```

Fig. 4. Favorability ratings from current Twitter data for current administration according to this system.

TABLE II
OUR SYSTEM WITH NO INFLUENTIAL CONSIDERATION.

	Debate 1	Debate 2	Debate 3
Donald Trump	41.62%	43.31%	40%
Hillary Clinton	58.38%	56.69%	60%

the higher number of Trump tweets but with lower average sentiment polarity compared to Clinton.Tweets are considered favorable for a candidate if the polarity is greater than .3. Tweets are considered unfavorable if the polarity is less than 0.



NBC News|SurveyMonkey Debate Reaction Poll among likely voters who watched or followed coverage

Fig. 5. NBC polls.

The results from our system without any influential consideration can be seen in Tab. II. With no influential considerations, the results from our system were quite close to those of NBC. Fig. 6 & 7 display results from our system by geolocation with influential considerations. With the incorporation of the influential model, it can be seen that the results favored Trump. This is because of the large number of dormant or silent Trump supporters. The influential model takes into consideration, the amount of followers that each Trump supporting Twitter account has. This outweighs that of the Clinton following.

The debate results reflected in figure 7, accounting the influence factor, suggest that Donald Trump was in fact performing significantly better than what the mainstream media portrayed in its polling. Tab. II and Fig. 6 reveal a stark contrast in the debate performances of each candidate. Based on the shocking results of the presidential election where Trump won many key swing states, the mainstream media should have likely placed more weight in the influence factor when predicting the final outcome. The results from the final sentiment model with influence consideration suggests that Trump was in fact winning before the election, aligning with the silent majority theory asserting that there was a large segment of pro-Trump voters who were reticent to publicly opine.

Another important political trend is revealed when comparing Fig. 6 and Fig. 7. Instinctively, one would posit that post-election sentiment would favor the winner of the election. However, as of the week of April 24, 2017, Trump sentiment has declined rather significantly as illustrated in Fig. 6 when compared to the pre-election data in Fig. 7. The poor favorability ratings reflected in Fig. 5 and Tab. II lend credence to this trend and is likely due to many unfavorable actions Trump has taken since he has taken office.

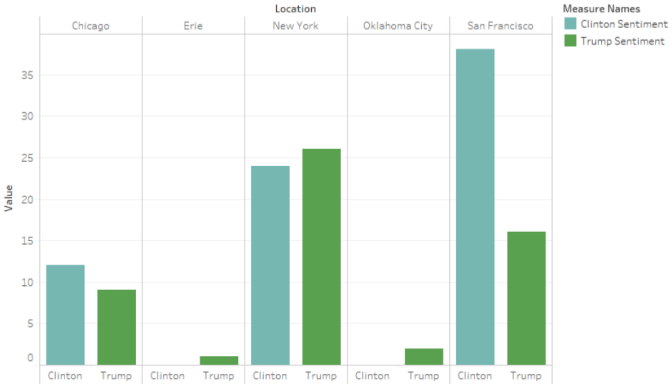


Fig. 6. Current Sentiment Data.

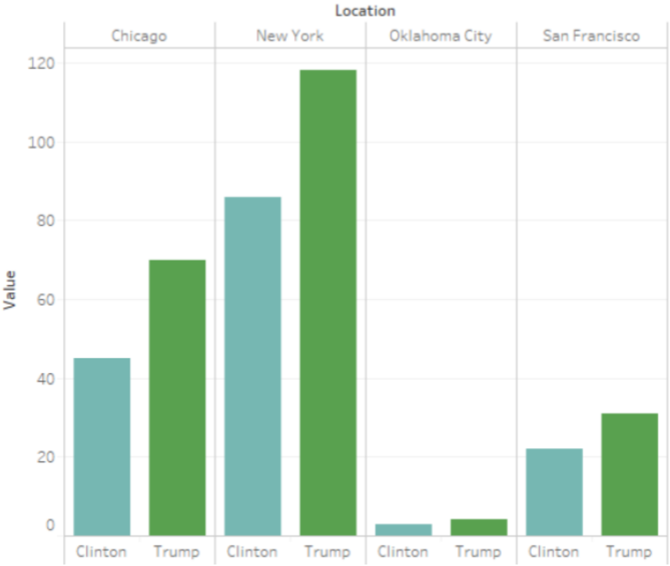


Fig. 7. Debate Sentiment Data

By calculating an influence value for each Twitter account, the key influencers for each side in debate 1 of the campaign can be seen in Fig. 8. This visualization gives an insight as to who the main influencers for each side are.

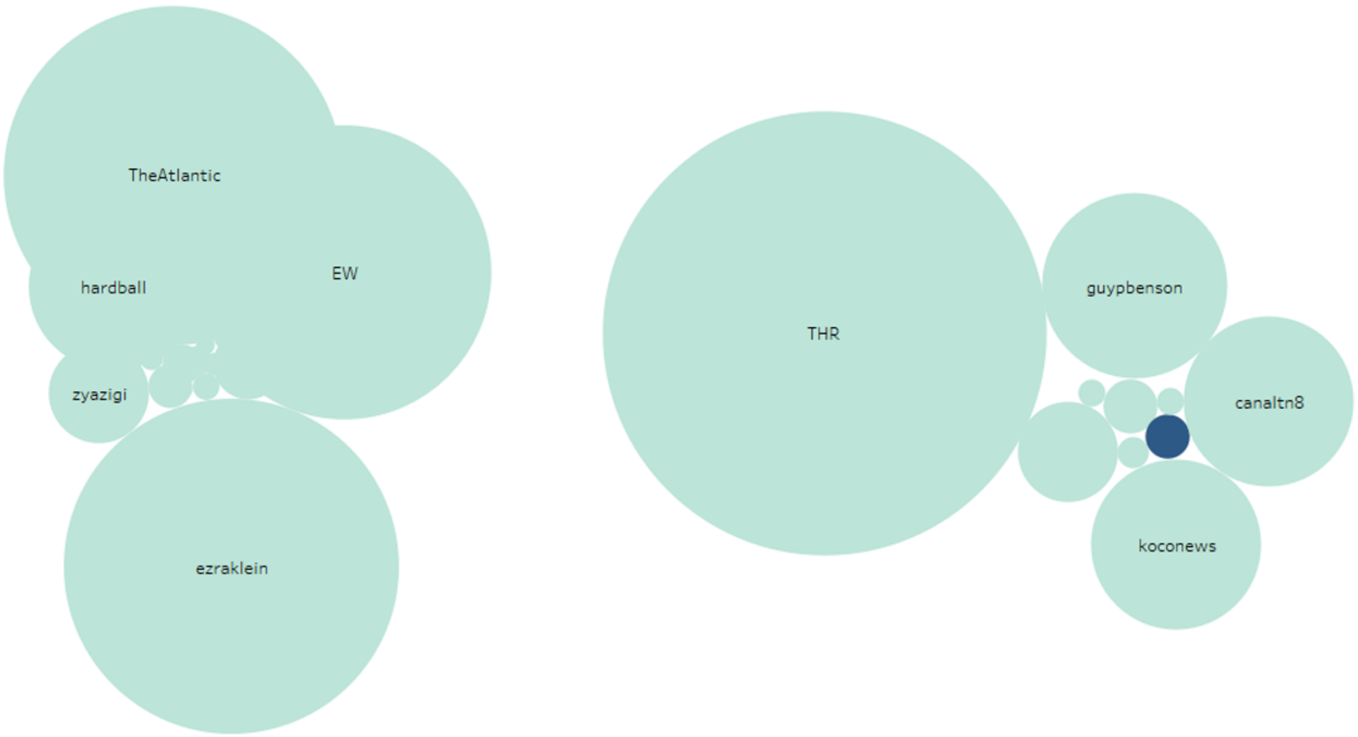


Fig. 8. Debate 1 Hillary (left) and Donald (right) key influencers.

TABLE III
DEBATE 1 SENTIMENT INSIGHTS.

Trump Sentiment	1,297
Influence Weighted Sentiment Trump	2,614
Clinton Sentiment	905
Influence Weighted Sentiment Clinton	8,346

Furthermore, the sentiment insights shown in Tab. III show how even with less tweets for a candidate, the candidate may still be more popular on Twitter as a result of our influence model. In this table, Trump received more positive tweets but Clinton had a greater influence. As silent Twitter users become more engaged throughout the campaign, this result shifts from being Clinton favorable to Trump favorable.

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

VI. LIMITATIONS

Many political tweets tend to be quite sarcastic. This makes sentiment analysis of these tweets very difficult.

It is very common for tweets to not contain geolocation data. As a result, geolocations 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 the 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 is an issue when calculating the influence of a user. Getting current Twitter user information is difficult. When calculating a detailed influence value like Klout [3], further Twitter user information is required.

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

VII. CONCLUSION

Twitter data reflected the outcome of each debate relatively accurately. New York tweets favored Trump, San Fran favored Clinton, Chicago was quite even. By tapping into inactive or lurking Twitter accounts, a more accurate prediction could be made for the outcome of the 2016 Presidential Election using Twitter data. Using this same model, Twitter sentiment shows that Donald Trump is faring very unfavorable by the populous at the end of his first 100 days.

VIII. 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 2016Ł Debate 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

[7] Title: Klout Score: Measuring Influence Across Multiple Social Networks, Author: Adithya Rao, Nemanja Spasojevic, Zhisheng Li and Trevor DSouza, URL: https://arxiv.org/pdf/1510.08487.pdf

[8] Title: Efficient Twitter Sentiment Classification using Subjective Distant Supervision, Author: Tapan Sahni, Chinmay Chandak, Naveen Reddy, Manish Singh, URL:https://arxiv.org/pdf/1701.03051.pdf

[9] Title: Measuring Influence on Twitter, Author: Isabel Anger, Christian Kittl, URL: http://www.l2f.inesc-id.pt/ fmmb/wiki/uploads/Work/misnis.ref07.pdf

[10] Title: Measuring User Influence in Twitter: The Million Follower Fallacy, Author: Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, Krishna P. Gummadi, URL: http://twitter.mpi-sws.org/icwsm2010_fallacy.pdf

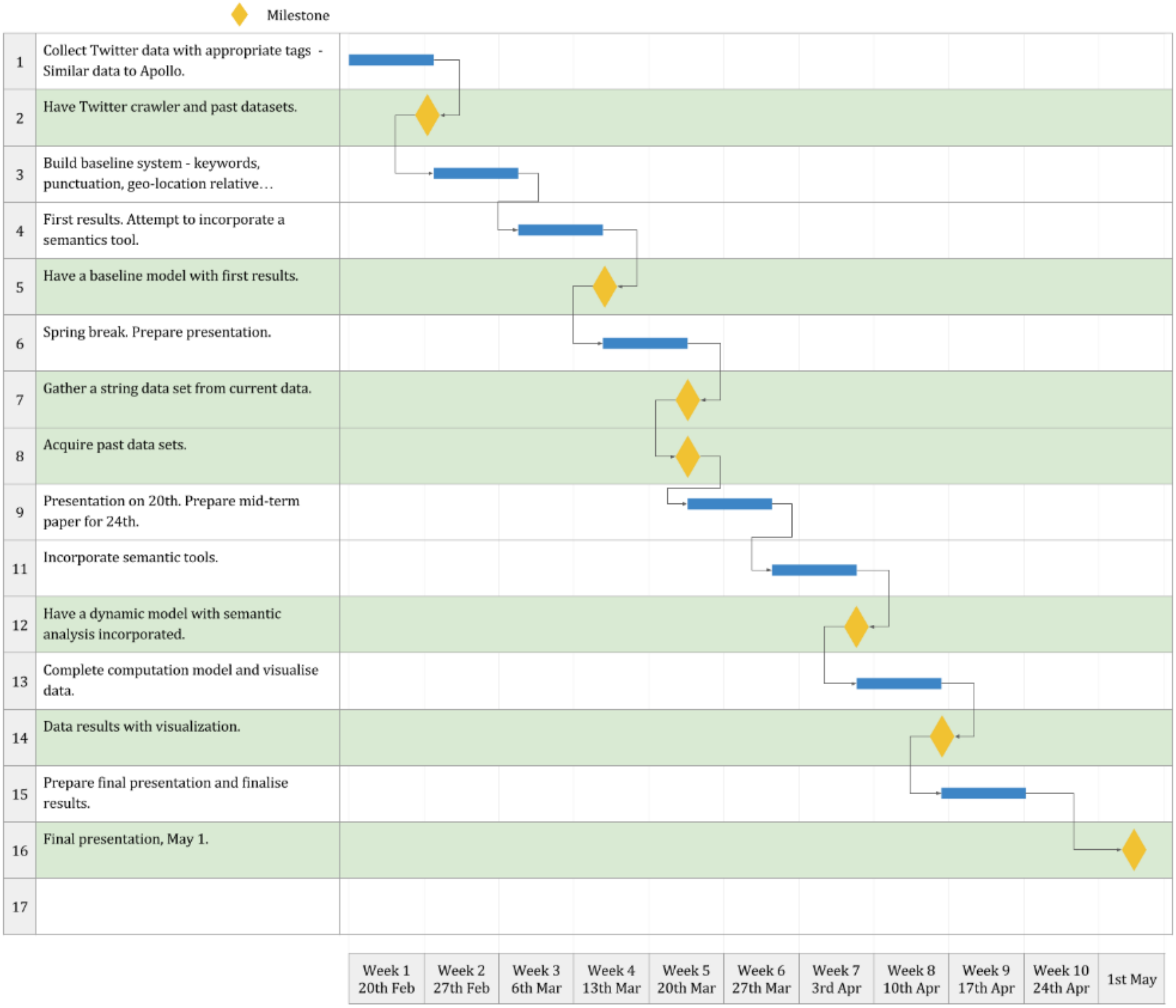


Fig. 9. Milestones Gantt chart.