Topic: Using Twitter Sentiment Analysis to Gauge Electability of U.S. Presidential Candidates

Your Name
Date

Background:

Presidential election results impact the direction of the entire country as well as foreign relations. For this reason there has always been a lot of interest in predicting the results of an election before it takes place. Traditionally, this job has been performed by pollsters, who call a sample registered voters, ask who they will vote for, and extrapolate the results for the entire voting population. Recently, with the help of machine learning techniques, researchers have looked to other mediums to predict election results due to the sometimes unreliable nature of traditional polls. For example, Véronis (2007) found that the number of candidate mentions in print journalism was a better indicator of election results than traditional polls. With the advent of social media, there has been a lot of research to determine whether user posts or statuses in aggregate can provide a prediction of election outcomes. There are two general methods to perform this type of analysis: use relative counts Tweets that mention each candidate to predict election outcomes or determine the relative sentiment of Tweets that mention each candidate to predict election outcomes. The results have been mixed with some research indicating that the mere relative count of Tweets can predict an election outcome, while other research has shown that Tweet counts and Tweet sentiment analyses are inaccurate and inconsistent. In his literature review of this topic, Gayo-Avello found that the problems with using Twitter to predict elections are (1) the user population does not represent the voting population as it is skewed towards younger people, and (2) Tweets are not 'cleansed' before being processed for analysis. Cleansing the Tweets means removing posted tweeted by users who are not part of the voting population, removing Tweets from political spammers, or removing Tweets that have nothing to do with the candidates' political platforms. As elections in the United States are overly covered by the media and tabloids, there is often focus on a candidate for reasons other than their politics. For example, Hillary Clinton's pantsuits or Donald Trump's hair is often a punchline. Removing these 'non-political' Tweets will provide a better basis for gauging electability of candidates.

Data Sources:

I plan to use Twitter data using the Twitter public REST API (https://dev.twitter/com/rest/public) and will be archiving the "Fire hose" real-time stream (https://dev.twitter.com/streaming/firehose) using Professor Brown's Twitter scraping script. All of the analysis I perform will be in R. I plan to archive tweets containing that mention seven candidates, five Republicans and two Democrats: Trump, Clinton, Bush, Fiorina, Rubio, Carson, and Sanders. I have chosen these names as these are the most commonly polled against candidates and these polls will serve as a benchmark to evaluate my results. I will limit my

Tweets to those that are geolocated in the United States in an attempt to exclude Tweets from individuals who are not eligible to vote. I will process the tweets in R using the text mining package "tm" (https://cran.r-project.org/web/packages/tm/vignettes/tm.pdf).

General Methodology:

The general method I will use in my project will be:

1. Collect Tweets and process using tm package
2. Use topic modelling to determine topics of Tweets for each candidate
    a. Split topics into 'political' and 'non-political'*
3. Use ensemble classification methods to classify Tweets into 'political' and 'non-political'
4. Perform sentiment analysis on all Tweets and 'political' Tweets only
5. Compare results for pairs of candidates to gauge relative electability
6. Compare Twitter results to polling results

*If no 'non-political' topics are found, then the project's focus will change to cleansing the Tweets by trying to remove political spam.

Algorithms:

Topic Modeling

Topic models are probabilistic algorithms that aim to identify topics or themes in a collection of documents. It is an unsupervised learning technique as the topics are previously unknown and are discovered by the algorithm. The topic modeling algorithm that I will be using is the Latent Dirichlet Allocation (LDA) algorithm. This is one of the most well-known topic models because it is the simplest. For this algorithm to work, each tweet must be converted into a sparse vector of 1s and 0s where each 1 indicates that a certain word is in the Tweet. The idea behind LDA is that the words in each document (in our case, Tweets) result from various topics, where each topic is a fixed multinomial distribution of words. LDA assumes that each document has been generated by a two-step process: (1) a distribution of topics is randomly chosen from all possible topics and (2) each word in the document is randomly chosen from the fixed distribution of the topics in the first step. Therefore LDA uses the observed words from a collection of documents (Tweets) to determine the underlying topic distributions in the documents and word distributions in the topics (Blei, 2012).

In R, LDA can be performed using the topicmodels package - https://cran.r-project.org/web/packages/topicmodels/vignettes/topicmodels.pdf

Supervised Classification:

Because it may be difficult to classify Tweets into politically intentioned Tweets and more humor/tabloid Tweets based on the results of the topic modeling, I have chosen to use ensemble classification methods rather than choosing just one algorithm in an attempt to get a better result. Ensemble classification methods use multiple learning algorithms to obtain a more accurate classification. In R, I will use the

RTextTools package (https://cran.r-project.org/web/packages/RTextTools/RTextTools.pdf), which is specifically designed for text classification. This package allows the user to train their data with up to nine algorithms: support vector machine, glmnet, maximum entropy, scaled linear discriminant analysis, bagging, boosting, random forest, neural networks, and regression tree. The package allows for easy comparison of the precision, recall, and F-score of each other algorithms on the test data in order to determine which combination of algorithms to use.

Sentiment Analysis:

While Tweet count methods and Sentiment Analysis have been used for research in this area, I will be using Sentiment Analysis as it has yielded better results for predicting election results according to Gayo-Avello. Sentiment analysis is more useful because much of the political traffic on Twitter is negative towards candidates and therefore merely counting Tweets could provide backwards results.

I will be using the "qdap" package in R for the Sentiment Analysis (https://cran.r-project.org/web/packages/qdap/qdap.pdf) . The ancillary package "qdapDictionaries" contains positive and negative word lists that were developed by M. Hu and B. Liu in 2004. I can add to these word lists if I determine that some of the Tweets I am processing are incorrectly classified in terms of sentiment. The 'polarity' function (http://www.inside-r.org/packages/cran/qdap/docs/polarity) in qdap calculates a polarity score by cross referencing the text of interest with the list of positive words with weight +1 and the list of negative words with weight -1. The polarity function also takes into account negating words such as "not" as well as amplifying and de-amplifying words such as "extra" and "barely".

Evaluation of Results:

The topic of my research is framed as "Using Twitter Sentiment Analysis to *Gauge Electability* of U.S. Presidential Candidates" rather than to *Predict Election Outcomes* because I will be collecting data from October –December 2015 during the primary campaigns. We will not have results of the primary elections nor the general election until this project is complete. Therefore, in order to evaluate the results of my analysis, I will compare my results to election polls, in particular, nation-wide "head-to-head" polls that consider democratic primary candidates versus republican candidates. These polls give a sense of the relative electability of candidates in general elections more than traditional primary polls that poll candidates of the same party against each other. The head-to-head polls ask the public the question, "If these two candidates were to win the primaries, which would you vote for". The idea behind this type of polling is to find the electability of a candidate at the general election level. For example, Donald Trump may out poll Jeb Bush among Republican voters, however, these polls may indicate that Jeb Bush will do better against Democratic candidates in the general election than Donald Trump.

The website Real Clear Politics, (http://www.realclearpolitics.com/epolls/latest_polls/president/), compiles results from various polls, including national polls for primary candidates from PPP, Quinnipiac, NBC/WSJ, and Fox. Because it is primary season in U.S. politics these polls take place every couple of

days. Below I have provided an example from October 6 to show what these polling results look like. Note that not all of the candidates get selected for head to head polling so I can only include the "Top Tier" primary candidates in my analysis. Because there is some bias towards certain parties in different polls, I will compare against as many polls as possible, including aggregated averages of polls over a time period.

General Election: Trump vs. Clinton	PPP (D)	Clinton 44, Trump 44
General Election: Bush vs. Clinton	PPP (D)	Bush 43, Clinton 42
General Election: Fiorina vs. Clinton	PPP (D)	Fiorina 43, Clinton 44
General Election: Rubio vs. Clinton	PPP (D)	Clinton 43, Rubio 43
General Election: Carson vs. Clinton	PPP (D)	Carson 48, Clinton 44
General Election: Trump vs. Biden	PPP (D)	Biden 48, Trump 40
General Election: Fiorina vs. Biden	PPP (D)	Biden 46, Fiorina 40
General Election: Rubio vs. Biden	PPP (D)	Biden 45, Rubio 40
General Election: Carson vs. Biden	PPP (D)	Biden 42, Carson 45
General Election: Trump vs. Sanders	PPP (D)	Sanders 44, Trump 44
General Election: Fiorina vs. Sanders	PPP (D)	Fiorina 44, Sanders 38
General Election: Rubio vs. Sanders	PPP (D)	Rubio 42, Sanders 38
General Election: Carson vs. Sanders	PPP (D)	Carson 46, Sanders 35

References:

Abramowitz, A. (1989). Viability, Electability, and Candidate Choice in a Presidential Primary Election: A Test of Competing Models. The Journal of Politics, 977-992. doi:10.2307/2131544.

Blei, D. (2012). Probabilistic topic models: Surveying a suite of algorithms that offer a solution to managing large document archives. Communications of the ACM Commun. ACM, 77-77.

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. Proceedings of the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '04, 168-177. doi:10.1145/1014052.1014073

Jahanbakhsh, K., & Moon, Y. (2014). The Predictive Power of Social Media: On the Predictability of U.S. Presidential Elections using Twitter. Computing Research Repository.

Tsakalidis, A., Papadopoulos, S., Cristea, A., & Kompatsiaris, Y. (2015). Predicting Elections for Multiple Countries Using Twitter and Polls. IEEE Intell. Syst. IEEE Intelligent Systems, 10-17.

Tumasjan, A., Sprenger, T., Sandner, P., & Welpe, I. (2010). Election Forecasts with Twitter - How 140 Characters Reflect the Political Landscape. Proc. 4th Int'l AAAI Conf. Weblogs and Social Media, 178–185.

Metaxas, P., Mustafaraj, E., & Gayo-Avello, D. (2011). How (Not) to Predict Elections. 2011 IEEE Third Int'l Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third Int'l Conference on Social Computing, 165-171.

Gayo-Avello, D. (2013). A Meta-Analysis of State-of-the-Art Electoral Prediction From Twitter Data. Social Science Computer Review, 31(6), 649-679.

Jurka T., Collingwood L., Boydstun A., Grossman E. & van Atteveldt, W. (2012). RTextTools: Automatic Text Classification via Supervised Learning. R package version 1.3.9.

Véronis, J. (n.d.). Citations dans la presse et résultats du premier tour de la présidentielle 2007. Retrieved from http://aixtal.blogspot.com/2007/04/2007-la-presse-fait-mieux-que-les.html