Predicting Party Affiliation Using Social Media

Joseph Brown, Mikaela Jordan, and Adam Swayze

Tarleton State University

Introduction

During primary elections, polls change quickly and are inaccurate. The polls from the days leading up to the Wisconsin Republican Primary showed changes in the winner within a day of the primary election. Finding a better way to predict primary elections is the goal of many political statisticians, and some have used social media to visualize which candidate the majority of a county supports.

The goal of the project is to predict which candidate will win a primary election in a county based on the words used in tweets originating in that county. Natural language processing is used to make the computer "understand" language. Word clouds can be used as a visualization for this goal. They depict the frequency of word use in a document by size and shading. The two word clouds shown here (Figures 2 and 3)come from twitter searches for the keywords "democrat" and "republican."

- "Bernie"
- "Sanders"
- "Feel the Bern"
- "Bernie2016"
- "I'm with Her"
- "HillaryClinton"
- "Hillary2016"
- "Clinton"
- "Trump"
- "Donald"
- "Make America Great Again"
- "Trump2016"
- "DonaldTrump"
- "Kasich"
- "JohnKasich"
- "Kasich2016"
- "Cruz"
- "TedCruz"
- "Trust Ted"
- "TrusTed"
- "Cruz2016"

Figure 1: Keywords for Twitter Search

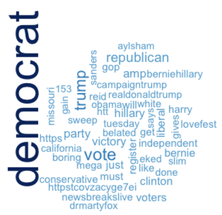

Figure 2: Word Cloud for "Democrat" Twitter Search

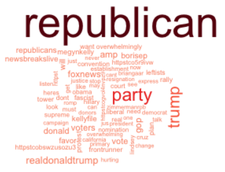

Figure 3: Word Cloud for "Republican" Twitter Search

Design of Primary Prediction

The methods of our project can be split into three categories: collecting data, cleaning tweets, and model design.

In order to collect the data, a dataset with the names of all counties in the U.S., the FIPS code for each county (unique identification number for each county in U.S), and the recorded majority winners for each county was created. A second dataset with the area and centroid of each county was found to set locale for the Twitter search. The Twitter search encompassed every county in 8 states (Arizona, Florida, Illinois, Indiana, Missouri, North Carolina, Ohio, and Utah). Twenty-one keywords were included in the search. (Figure 1)

After the Twitter search completed, the dataset with the results from each county was merged with tweets data frame by FIPS code.

When tweets are collected, the resulting document has many impurities in the text. For ease of analysis, the text is cleaned. Since the project is based off of term frequency, punctuation is removed and all words are forced into lower case to decrease duplicate terms. Emojis and stop words (the most common words in the English language) are removed from the text because they will not affect the analysis, and will only make the document term matrices larger.

For the models, a document-term matrix (DTM) was created from all of the tweets, and a label was attached to each tweet based on the location of the tweet (and which candidate won the Republican and Democrat primary in that county). Once the DTM was labeled, machine learning algorithms (Support Vector Machine, Artificial Neural Network, and Random Forest) were applied to the data to predict the majority winners of each tweet's county.

Democrat Results

The highest accuracy rates for the democratic party primary predictions were associated with a support vector machine learning algorithm and a neural network with size 6 model.

Democrat Results Cont'd

		Predicted	
		Bernie	Hillary
Actual	Bernie	431	3332
	Hillary	148	10212

Table 1: SVM Confusion Matrix

		Predicted	
		Bernie	Hillary
Actual	Bernie	434	3329
	Hillary	132	10228

Table 2: Neural Network Confusion Matrix

	Accuracy Rate
Support Vector Machine	75.35934%
Neural Network	75.49388%

Table 3: Accuracy Rates of Both Models

Tables 1 and 2 are confusion matrices, and table 3 contains the accuracy rates of both models. The neural network model has a 0.13% accuracy rate than the support vector machine model.

Republican Results

There are three candidates for the Republican primaries and Trump is winning a far greater number of counties than Cruz and Kasich combined. Due to those constraints, the models are predicting "Trump" or "not Trump" rather than each candidate individually.

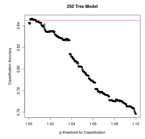

Figure 4: Classification accuracy vs Probability threshold

		Predicted	
		Trump	Not Trump
Actual	Trump	9517	42
	Not Trump	1695	58

Table 4: Random Forest 250 Trees Confusion Matrix

This peak in the graph at .976 shows that in order for the random forest model to predict that a county would go "not Trump," there had to be 97.6% certainty in that claim (typically this value is set at 50%). The model has an accuracy of 84.6%, only slightly higher than if "Trump" was predicted for all counties (84.5%).

Limitations

There is a class imbalance issue, as Donald Trump is wining many more precincts than the other Republican candidates. This skews the model in favor of Trump. On the Democratic side, a majority of Bernie Sanders supporters are younger and more likely to use Twitter. According to the Pew Research Center, 88% of people 21 and younger are active on Twitter. Some limitations come with the API used to collect tweets, and restrictions from Twitter. Only 23% of Americans use Twitter, so the sample might not be representative of the entire population. Of the tweets that are available, less than 5% of tweets are georeferenced (which is necessary to be able to label the tweets to train the models).

Future Work

The models are continuing to be updated as more primary elections occur. In the future, the models will be improved to account for other geographic differences such as socioeconomic status of precincts. Debate transcripts will be used in order to predict partiality of media networks using online articles. The ultimate goal of this project is to extend and improve the current models to predict the outcome of the presidential election this November.

References

Benyamin, Dan."A Gentle Introduction to Random Forests, Ensembles, and Performance Metrics in a Commercial System - Blog & Press" *Citizen and Net.* N.p., 09 Nov. 2012. Web. 28 Mar. 2016.

Phillips, Winfred. "Introduction to Natural Language Processing." *Consortium on Cognitive Science Instruction.* The MIND Project, 2006. Web. 16 Mar 2016.

Raine, Lee. "Social Media and Voting." *Pew Research Center Internet Science Tech RSS.* Pew Charitable Trusts, 05 Nov. 2012. Web. 28 Mar. 2016.

Weston, Jason. "Support Vector Machines (and Statistical Learning Theory)." ComputerScience4701. Columbia University, New York. *Columbia University.* Web. 28 Mar. 2016.

"FiveThirtyEight." *FiveThirtyEight.* Nate Silver, n.d. Web. 28 Mar 2016.

"2016 Primary Election Results: President Live Map by State, Real-Time Voting Updates" *Election Hub.* Politico LLC, 27 Mar. 2016. Web. 28 Mar. 2016.