**Applying Topic & Sentiment Analysis in the Political Arena**

**Introduction**

The United States held elections for 435 House and 35 Senate Congressional seats in 2018. Typically, there would have been 34 Senate elections in 2018; however, Minnesota had a resignation of one of its Senator’s mid-term and a ‘special election’ was held in 2018 for this seat. Election cycles have become increasingly partisan, and this has been spilling into the various Congressional races. How the increased partisanship has impacted these races has varied a bit, but what has not changed is that policy positions continue to be an important determinant of a candidate’s success. With the shifting demographic, partisan and ideological behaviors of voters, the difficulty for candidates of determining ideal policy positions has increased.

Polling is the primary method by which campaigns typically determine which issues matter and just how much they matter. Polling has also become progressively more difficult and expensive, and is subject to issues like missing turnout estimates, non-response bias and difficulty in wording questions in such a way as to not unduly bias responses. Indeed, pollsters typically need to call 65 people for every 1 person they get to respond by phone. Polls conducted online are subject to a whole host of other challenges.

This project was aimed at demonstrating that natural language processing techniques could be used to (a) add insight as to which issues matter most in those electoral campaigns and (b) that those identified issues could include a sentiment score for each candidate, informing an ideal policy position for those candidates. If successful, such techniques could be used in lieu of polling, or at the very least minimize how much polling is required.

**Research Design & Methods**

The scope of this project did not allow for covering all 470 Congressional races in 2018. Rather, a sampling of the most competitive races was used as a ‘proof of concept’ starting point. Key criteria for picking races included (a) sufficient data to run analysis and (b) meaningfully high probability of the race being competitive. House races, given how much smaller they typically are compared to Senate races, were dismissed from consideration immediately. Of the 35 Senate races remaining, most were not competitive. From the remaining races, Arizona, Florida and Missouri Senate races were chosen as races being sufficiently large and competitive.

A. Data

Data sources used for this project included news articles and user “tweets” from Twitter, Inc. These data sources offered some measure of internal validation, as producers of twitter and news media do not overlap too much. Both data sources offered their own set of challenges though.

Twitter data in particular was challenging to (a) collect, (b) transform into a useable format, and (c) adjust for retweets. For these reasons, RapidMiner, a commercial data science software platform, was used to pull tweets directly from Twitter. Twitter throttles attempts to scrape it’s data, so it was necessary to run a collection algorithm dozens of times to obtain a complete dataset. 14% of tweets were cut off, but did include a link to the full tweet. The researchers dealt with this by programmatically looping over all tweets with a link matching the appropriate link format to retrieve full tweet text. Retweets were removed, although the number of retweets was recorded in our dataset for further analysis. Unsurprisingly, more than 80% of the collected tweets were generated from less than 20% of the Twitter users in the sample set. In all, a total of about 160,000 tweets were collected between October 16 and October 23, 2018 for the six Senate candidates.

Obtaining news data was a different challenge. There was a need to obtain news articles related to the chosen candidates; however, not all news outlets covered those candidates. As a result, identifying news websites that focused their political articles on these candidates became the mission. Using Google to search for the candidates helped identify numerous news outlets, most of which were local to the cities and states where the candidates resided.

Once these websites were identified, Python code was written that utilized the *requests*, *BeautifulSoup*, and *newspaper3k* Python libraries to perform web scraping against all these news websites. Once per day for the month of October, the Politics section of each of these news websites [Appendix 1] were scraped looking for all articles where the candidates were mentioned. At the end of the month 144 unique news articles had been collected for analysis.

B. Topic Selection

In order to obtain the primary topics for each tweet and news article, another Python script was constructed. Each tweet and article were tokenized allowing the use of the *gensim* python library within this script to construct a Latent Dirichlet Allocation (LDA) model. This LDA model was able to provide keywords from each tweet and each news article, which were then used to identify topics for each. The initial expectation was to identify political platforms of candidates (for instance: gun control, immigration, and healthcare); instead, LDA returned popular words that would require manual translation to form platforms..

C. Sentiment Analysis

A sentiment score was allocated to each tweet and news article in order to obtain an idea of how the author of the text felt about the candidate and the specific topic. Using the *Afinn* Python library, sentiment scores were obtained and the scores ranged from -56 to +93. Negative scores reflect poor feelings towards the the topic being discussed and positive scores indicate positive feelings towards the topic being discussed. There was one outlier (+268) that was removed from the analysis.

**Results**

Running the topic and sentiment analysis in Python created a comma separated values (CSV) data file that contained output values for each article and tweet. Tableau Desktop was then used to explore the results and create visualizations [Appendix 2]. Within these visualizations it can be seen that sentiment score by source showed a wider spread on average for news (-10.4 to 18.8) than tweets (-3.2 to 3.1) [2a]. The individual score distributions for tweets hug the origin more, reinforcing a tighter range [2b,c].

Sentiment score by candidate was then further broken down by winning and losing campaign races. The winning candidates had an average score of 19 score while the losing candidates had an average score of 10 [2d]. Tableau’s group function was used to aggregate articles by news company. These results showed in general, a newspaper either had a positive sentiment for both opposing party candidates or a negative sentiment for both [2f]. The top positive and negative sentiment scores by topics by candidate were visualized on a heatmap. The +93 sentiment score datum was about Rick Scott and had the topics of *scott*+*florida+nelson+say+trump* [2g]. The worst scored datum was about Josh Hawley and about *polit+campaign+hawley+general+offici* [2h]. Finally, a word cloud was generated to display topic analysis results. The frequency of topic is represented by size while the sentiment score associated with the article or tweet represented by color. The topic *scottforflorida* was the most popular topic, and was associated with positive sentiments [2i].

**Discussion**

Topic selection seems to work well, but there is definitely some additional work to be performed to translate the word groupings into issues that are more readily identifiable in a political campaigning environment. Key issues that were identified, and do seem to matter from these races, include healthcare, trump, and a bombing attempt made on Democratic leaders toward the end of the 2018 campaign.

News also appears to be a more positive source of news than twitter. Not all that surprising given Twitter’s reputation for being quite negative, although what is surprising is the analysis also shows Twitter to be roughly equal in terms of its positive and negative sentiment. There also seems to be a loose correlation between campaigns that are, on balance, more positive and winning their elections, although it’s difficult to make any real determinations with only six campaigns from which to draw conclusions.

As aforementioned, this project was a ‘proof of concept.’ From that perspective, the project was successful. Possible extensions of this approach could be to scale it up, looking for any complications as more campaigns are included. Additional data sources could also be obtained, which would aid in the low sample sizes. A more automated, or ‘awareness algorithm’ could be triggered when a new relevant news article or tweet was published. The model and approach could be integrated with an election result forecasting platform. Overall positive sentiment does seem to be associated with a campaign winning its election; however, is this observation statistically valid and scalable? Can the topic selection process be developed to map onto more typical campaign issues?

The possibilities for an approach like this could also be extended outside the political arena to general markets, focusing on brand management or policy.

**Appendix**

[1] List of news websites that articles were obtained from (.json format):

{

"apnews\_politics": {

"link": "https://www.apnews.com/apf-politics" },

"florida\_today\_politics": {

"link": "https://www.floridatoday.com/section/global/elections" },

"tampabay\_politics": {

"link": "https://www.tampabay.com/florida-politics/" },

"sun\_sentinel\_politics": {

"link": "https://www.sun-sentinel.com/news/politics/" },

"nevada\_indep\_elections": {

"link": "https://thenevadaindependent.com/election/2018/governor" },

"lasvegas\_now\_elections": {

"link": "https://www.lasvegasnow.com/your-local-election-headquarters" },

"lasvegas\_now\_politics": {

"link": "https://www.lasvegasnow.com/news/politics" },

"lasvegas\_sun\_politics": {

"link": "https://lasvegassun.com/news/politics/" },

"reno\_election": {

"link": "https://foxreno.com/news/election" },

"stlouis\_politics": {

"link": "https://www.stltoday.com/news/local/govt-and-politics/" },

"kansas\_city\_politics": {

"link": "https://www.kansascity.com/news/politics-government/" },

"kansas\_city\_politics": {

"link": "https://www.kansascity.com/news/politics-government/election/" },

"vox\_politics": {

"link": "https://www.vox.com/policy-and-politics" },

"foxnews\_politics": {

"link": "http://www.foxnews.com/politics" },

"cnbc\_politics": {

"link": "https://www.cnbc.com/politics/" },

"washingtonpost\_politics": {

"link": "https://www.washingtonpost.com/politics/" },

"onion\_politics": {

"link": "https://politics.theonion.com/" },

"cnn\_politics": {

"link": "https://www.cnn.com/politics" },

"bbc\_us\_politics": {

"link": "https://www.bbc.com/news/topics/cwnpxwzd269t/us-politics" }

}

[2] Public Tableau of results available at:

<https://public.tableau.com/profile/akash.jassal#!/vizhome/TopicAndSentiAnalysis_DATS6401_NLP_Presentation/Presentation_1>



