Exploring the relationship between press coverage and polling performance

I explore the relationship between press coverage and poll results by studying the 2016 presidential election. The main aim of this project to learn how to deal with GDELT data and make a basic US election prediction model.

The approach is to build a data set consists of poll results and key press coverage metrics and tones between June 2016 and October 2016. Using these data, we can then explore relationships.

Data

To conduct this analysis both polling and press-coverage data are required. These were collected from two different sources, FiveThirtyEight, and GDELT project.

GDELT is supported by Google Jigsaw and is a project that’s monitors, stores and provides the world’s broadcast, print and web news.

For this project, data were pulled from the Global Knowledge Graph Version 2 (GKG V2) database. The GDELT GKG values for document counts, tone, positive and negative scores, polarity, activity reference density and self-reference density and a compilation of article themes were captured for articles related to the primary candidates. More information on these scores can be found in the GDELT

[documentation](http://data.gdeltproject.org/documentation/GDELT-Global_Knowledge_Graph_Codebook-V2.1.pdf)

The GKG V2 database is available as a public dataset accessible at BigQuery. However, it is quite large at over 1.2 terabytes (TB). This size requires that queries are carefully constructed to minimize running time and importantly to minimize cost. While Google makes a certain level of searching available for free, BigQuery is a fee-based service that charges for both storage and the data consumption for queries. Repeated queries of the full CKG V2 database will generate fees which can be costly.

Process

We access GDELT data without Google Bigquery (pip install gdelt) and use Google Colab. GDELT collects the news data every 15 minutes. Average daily data size is about 1.5GB. It was a huge size of data to work on the Google colab. Therefore, we only use daily 11:45PM data to save RAM storage. We consider 2 main candidates for the analysis; republicans and democrats, collect the data for president and vice president candidates. Trump and Pence for republicans and Clinton and Kaine for the democrats.

First, we get the count of news based on the candidates’ name, then calculate average tone rate in daily bases by each state. Tone is calculated as the difference between positive and negative scores. Now, we have a daily basis each candidate’s count of news and average tones information by each state.

Second, we get the poll data from June 2016 to October 2016 for the same candidates by each state. Now, we can find a relationship between GDELT and poll data. But, we also assume that news data may not affect the poll data at the same day. Therefore, we calculate the number of news and average of the tones by week. Then, we do the same process for the poll data and calculate the average estimation poll data per week. Then, we combine GDELT with poll data.

Third, we run Xgboost implementation to train our data for 2016.

Finally, we collect the same GDELT data for 2020 from June to September. With this data, we use our model to predict 2020 poll results.