Unsupervised Learning

Identifying Political Platforms from Twitter data using NLP

ABSTRACT

Party platforms were once thought to be essential to the electoral process as they give the candidates a clear political position with which they can campaign. These platforms give voters a sense of what the candidates believe in, the issues they will focus on and if elected, how they will address them in their policy-making. This project used NLP methods and tools to identify the possible political platforms for the 2022 US Midterm Elections.

DESIGN

Using Twitter which is the modern political soapbox, this project aims to use tweets to identify the Democratic and Republican political platforms by analyzing tweets for topic patterns. Tweets and topics with the most 'Reach[1]' and their sentiment reflects how the twitter-active voting public might respond to a political agenda. The top 7 'Topics' for each political party based on this data could be the political platforms or planks for the upcoming election.

TOOLS

- Pandas, Numpy, re, string and Matplotlib for EDA
- VADER, FLAIR and TextBlob for Sentiment Analysis
- nltk, spaCy, sklearn for Modeling

ALGORITHMS

- Feature Engineering:
    - Retweet and Favorite columns were scaled and aggregated to create 'Reach'
    - Sentiment Scores from FLAIR, TextBlob and VADER were averaged
- Sentiment Analysis:
    - TextBlob results were polarity and subjectivity of tweets
    - FLAIR gave a label score which shows likelihood of label
    - VADER gave sentiment scores and corresponding probability

[1] Reach is a feature created from the aggregation of Retweet and Favorites data from each tweet

- Topic Modeling:

    - Latent Semantic Analysis (LSA) , Latent Dirichlet Allocation(LDA) and Non-negative Matrix Matrix Factorization(NMF) were used before settling with NMF as the topic modeler due to the better interpretability of the results.
    - Given the number of tweets, the parameters that produced the best results were:
        - TfidfVectorizer
            - max_df = .95
            - min_df = 25
            - ngram_range=(1,2)
        - NMF
            - n_components=7
            - init='nndsvd'
    - For Democrats here are the topics and how their tweets rank:

NMF_Topic_name	Polarity	Compound	Reach
Covid	0.170293	0.236002	127055.5
Voting Legislation	0.108495	0.123206	68327.5
American Rescue Plan	0.096555	0.181038	34636.5
Infrastructure	0.142306	0.239092	20022.5
Healthcare	0.118381	0.276057	18691.0
Biden Presidency	0.091025	0.067997	15144.5
Build Back Better Act	0.195808	0.381086	9433.0

- For Republicans here are the topics and how their tweets rank:

NMF_Topic_name	Polarity	Compound	Reach
Covid	0.168992	0.241958	88237.5
Biden Presidency	0.010175	-0.130169	52831.0
Voting Legislation	0.095111	0.101531	41354.5
American Rescue Plan	0.059853	0.046947	24383.5
Infrastructure	0.108951	0.159856	10651.0
Build Back Better Act	0.151285	0.194882	5067.5
Healthcare	0.098707	0.213287	4694.0

CONCLUSION

Both Democrats would benefit from talking about Covid and the government response to it. The reach of tweets related to the topic seem overwhelmingly popular even though the quantity of tweets related to them are relatively lower than topics like Voting Legislation. Compound scores and Polarity for the Democrats tweets are more positive than Republicans.

Just like in 2020 it does seem that tweets about Biden and the administration are very important to Republicans, and being the only topic with a compound negative score, it shows that they can continue to focus less on important issues like health care or infrastructure since this is what their audience is expecting.

For Democrats, their tweets and how they are received are geared more toward policy and helping the American people and less on the presidency itself.

COMMUNICATION

This project and accompanying code can be found on github here.
A Streamlit app is currently under development to show the results of the project.

Submitted by: Mike Bernardo