Lakshmi S. Menon, Parthvi Sanjay Shah, Anusha R. Patil, and Muhammad Shujaat Mirza

COVID19: War of Twitter Narratives

Motivation

We are interested in the analysis of political discourse on Twitter in the US with respect to the ongoing `COVID-19` pandemic. The motivation behind this analysis is to identify political narratives that harnessed more support among the users of the micro-blogging site and that were pushed by different news agencies owing to their implicit political bias. To build the training corpus, we will utilize tweets pertaining to `COVID-19` made by the US Governors hailing from Democratic or Republican parties, allowing us an efficient way to identify and label tweets representing discourse pushed from either side of the political spectrum to deal with the same crisis. Using the trained model, we intend to analyze the Twitter accounts of all major US news outlets to identify the support for either of these political discourses and whether their stance evolved over time. For possible future work, another application of the model could be used to track real-time twitter sentiments in response to real-world events such as, how do the postings on the platform evolve following events, such as the White House briefings.

Proposed Approach

This is a supervised learning classification task, where each instance represents the tweet by a particular governor and the label is either `‘Republican’` or `‘Democrat’`. We will start by making a database of tweets published by governors from all 50 states. Tweets relating to coronavirus will be determined by searching for keywords and hashtags pertaining to the virus, such as `‘coronavirus’`, `‘COVID-19’`, `‘pandemic’`, `‘lockdown’`, etc. We will consider tweets within a fixed time range, starting from the date of the first reported U.S. case. We will use a combination of the `COVID-19` tweets datasets [2, 3] as well as tweets that we scrape using the `Twitter API` [1]. We will also need to assign party labels by scraping the name, state, and party from Wikipedia. To allow comparison between multiple models and hyperparamters, we will split the database into training, validation and test sets with the following ratios: `70% train, 15% validation, 15% test`. We will then use the best performing model to classify the tweets obtained from a separate database constituting the tweets from major US news outlets.

Evaluation Approach

Before training the model, we will use preprocessing to simplify our dataset and take out extra features that may add noise. In addition to common techniques such as removing punctuation, numbers, and stopwords, we will also need to do some preprocessing specific to our tweet data. This includes removing URLs, usernames, and media, as well as perhaps mapping emojis to a word representing the conveyed emotion. We would start out with a baseline model using a `bag-of-words` model, and then extend it to produce better results, using `n-grams`, `TF-IDF weighting, and naive bayes` model. We can also extend this to `logistic regression` and `random forest` models. Apart from doing sentiment analysis, we also plan to do topic modelling, to see if a particular party focuses more or less on a specific aspect of the coronavirus impact. This would be an unsupervised task, which we would perform using the `latent dirichlet allocation (LDA)` algorithm. The metrics we use to evaluate our models will be decided based on the properties of our dataset, such as class imbalance, as well as the algorithm used, but we will consider metrics such as accuracy, `auROC` and average precision for our classification models and coherence score for our `LDA` model.

References

[1] Twitter developer api. [Online]. Available: https://developer.twitter.com/en/docs

[2] R. Lamsal, “Corona virus (covid-19) tweets dataset,” 2020. [Online]. Available: http://dx.doi.org/10.21227/781w-ef42

[3] C. Lopez, “Covid-19 tweets dataset,” 2020. [Online]. Available: https://www.kaggle.com/lopezbec/covid19-tweets-dataset

Lakshmi S. Menon: lsm454; Responsible for submissions
Parthvi Sanjay Shah: pss434@nyu.edu
Anusha R. Patil: arp624@nyu.edu
Muhammad Shujaat Mirza: msm622@nyu.edu