Using Coronavirus Tweets to Predict Political Party

Group Members:

- Lakshmi S. Menon lsm454 (Member responsible for uploading submissions)
- Parthvi Sanjay Shah pss434
- Anusha R. Patil arp624

Summary of Plans

I. Project Topic

We have chosen to do a project related to the current COVID-19 pandemic. We will be analyzing tweets pertaining to coronavirus made by U.S. Governors to see if we can predict their political party. The main motivation behind this idea is to see if there is a difference in how members of different political parties respond to the same crisis. We may also apply our model to news articles to see if different news channels or sites have an implicit political bias.

II. Proposed Approach

This is a supervised learning binary classification task, where each instance represents the tweets by a particular governor and the label is either 'Republican' or 'Democrat'. Our dataset will be composed of tweets 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 Tweets Dataset from IEEE[1] as well as tweets that we scrape using the Twitter API. We will also need to assign party labels by scraping the name, state, and party from Wikipedia.

III. Suggested Experiments

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.

Reference:

[1] https://ieee-dataport.org/open-access/corona-virus-covid-19-tweets-dataset