Project 5
Natural Language Processing
Sociology 273M: Computational Social Science

This lab is divided into two parts. Each part has a different deadline.

Part I

1 Introduction

In this project, you will use natural language processing techniques to explore a dataset containing tweets from members of the 116th United States Congress that met from January 3, 2019 to January 2, 2021. The dataset has also been cleaned to contain information about each legislator. Concretely, you will do the following:

- Preprocess the text of legislators' tweets
- Conduct Exploratory Data Analysis of the text
- Use sentiment analysis to explore differences between legislators' tweets
- Featurize text with manual feature engineering, frequency-based, and vector-based techniques
- Predict legislators' political parties and whether they are a Senator or Representative

You will explore two questions that relate to two central findings in political science and examine how they relate to the text of legislators' tweets. First, political scientists have argued that U.S. politics is currently highly polarized relative to other periods in American history, but also that the polarization is asymmetric. Historically, there were several conservative Democrats (i.e. "blue dog Democrats") and liberal Republicans (i.e. "Rockefeller Republicans"), as measured by popular measurement tools like DW-NOMINATE. However, in the last few years, there are few if any examples of any Democrat in Congress being further to the right than any Republican and vice versa. At the same time, scholars have argued that this polarization is mostly a function of the

Republican party moving further right than the Democratic party has moved left.

Q1: Does this sort of asymmetric polarization show up in how politicians communicate to their constituents through tweets?

Second, the U.S. Congress is a bicameral legislature, and there has long been debate about partisanship in the Senate versus the House. The House of Representatives is apportioned by population and all members serve two year terms. In the Senate, each state receives two Senators and each Senator serves a term of six years. For a variety of reasons (smaller chamber size, more insulation from the voters, rules and norms like the filibuster, etc.), the Senate has been argued to be the "cooling saucer" of Congress in that it is more bipartisan and moderate than the House.

Q2: Does the theory that the Senate is more moderate have support in Senators' tweets?

2 Data

The dataset has been cleaned, and contains the following columns:

- **tweet_id**: ID Number for tweet
- **screen_name**: Legislator's Twitter handle
- **datetime**: Date and time tweet was posted
- **text**: Text of tweet
- **name_wikipedia**: Legislator's name as per Wikipedia
- **position**: Whether the legislator is a member of the U.S. Senate or House of Representatives
- **joined_congress_date**: The date that the legislator joined Congress
- **birthday**: The legislator's birthday
- **gender**: The legislator's gender
- **state**: The legislator's state
- **district_number**: For House members, their House district number. '0' indicates at-large district, 'Senate' indicates that the individual is a Senator
- **party**: The legislator's political party (usually Democratic or Republican)

- **trump_2016_state_share**: How many votes Donald Trump won in the state in the 2016 presidential election
- **clinton_2016_state_share**: How many votes Hillary Clinton won in the state in the 2016 presidential election
- **obama_2012_state_share**: How many votes Barack Obama won in the state in the 2012 presidential election
- **romney_2012_state_share**: How many votes Mitt Romney won in the state in the 2016 presidential election

Note that although there may be 535 members voting members of Congress at any time (435 in the House of Representatives and 100 in the Senate), we only have about 450 unique legislators in the dataset. Reasons for missingness include the legislator not having a twitter handle, not having certain demographic information available, or they did not complete a term due to death or resignation. Legislators who joined Congress after January 3, 2019 may also not be included. For instance, Mark Kelly (D-AZ) defeated incumbent Martha McSally (R-AZ) in the November 2020 election, but was sworn in on November 30, 2020 as part of the 116th Congress instead of January 3, 2021 as part of the 117th because that race was a special election to fill John McCain's (R-AZ) seat.

As you're working, keep a few things in mind:

- There are two independent Senators, Bernie Sanders (I-VT) and Angus King (I-ME), who both caucus with the Democratic Party.
- There were three party switches in the House during the 116th Congress. Justin Amash (L-MI) switched from Republican to Independent to Libertarian, Paul Mitchell (I-MI) switched from Republican to Independent, and Jeff Van Drew (R-NJ) switched from Democratic to Republican.
- There are some representatives who have a district number of "0". These states (Alaska, Delaware, Montana, North Dakota, South Dakota, Vermont, and Wyoming) have one "at-large" district that represents the whole state in the House of Representatives.
- There may be some inconsistencies that you find throughout the dataset. For instance, Kamala Harris' twitter handle was "@SenKamalaHarris" but after her election as Vice President, her twitter handles are "@KamalaHarris" and "@VP" so this was fixed. Wherever you find inconsistencies and are unsure how to resolve them, document the assumptions you make before proceeding with your analysis.

Overall, there are close to 1 million individual tweets in this corpus. Remember to start with smaller subsets of the data to make sure your code works before proceeding with a larger analysis! The dataset was constructed from a few different sources and required a substantial amount of preprocessing, so there may be some bugs or mistakes.

3 Text Preprocessing

The first step in working with text data is to preprocess it. Make sure you do the following:

- Remove punctuation and stop words. The 'rem_punc_stop()' function we used in lab is provided to you but you should feel free to edit it as necessary for other steps
- Remove tokens that occur frequently in tweets, but may not be helpful for downstream classification. For instance, many tweets contain a flag for retweeting, or share a URL

As you search online, you might run into solutions that rely on regular expressions. You are free to use these, but you should also be able to preprocess using the techniques we covered in lab. Specifically, we encourage you to use spaCy's token attributes and string methods to do some of this text preprocessing.

Part II

4 Exploratory Data Analysis

Use two of the techniques we covered in lab (or other techniques outside of lab!) to explore the text of the tweets. You should construct these visualizations with an eye toward the eventual classification tasks: (1) predicting the legislator's political party based on the text of their tweet, and (2) predicting whether the legislator is a Senator or Representative. As a reminder, in lab we covered word frequencies, word clouds, word/character counts, scattertext, and topic modeling as possible exploration tools.

5 Sentiment Analysis

Next, let's analyze the sentiments contained within the tweets. You may use TextBlob or another library for these tasks. Do the following:

- Choose two legislators, one who you think will be more liberal and one who you think will be more conservative, and analyze their sentiment and/or subjectivity scores per tweet. For instance, you might do two scatterplots that plot each legislator's sentiment against their subjectivity, or two density plots for their sentiments. Do the scores match what you thought?
- Plot two more visualizations like the ones you chose in the first part, but do them to compare (1) Democrats v. Republicans and (2) Senators v. Representatives

6 Featurization

Before going to classification, explore different featurization techniques. Create three dataframes or arrays to represent your text features, specifically:

- Features engineered from your previous analysis. For example, word counts, sentiment scores, topic model etc.
- A term frequency-inverse document frequency matrix.
- An embedding-based featurization (like a document averaged word2vec)

In the next section, you will experiment with each of these featurization techniques to see which one produces the best classifications.

7 Classification

Either use cross-validation or partition your data with training/validation/test sets for this section. Do the following:

- Choose a supervised learning algorithm such as logistic regression, random forest etc.
- Train six models. For each of the three dataframes you created in the featurization part, train one model to predict whether the author of the tweet is a Democrat or Republican, and a second model to predict whether the author is a Senator or Representative.
- Report the accuracy and other relevant metrics for each of these six models.
- Choose the featurization technique associated with your best model. Combine those text features with non-text features. Train two more models: (1) A supervised learning algorithm that uses just the non-text features and (2) a supervised learning algorithm that combines text and non-text features. Report accuracy and other relevant metrics.

If time permits, you are encouraged to use hyperparameter tuning or AutoML techniques like TPOT, but are not explicitly required to do so.

8 Discussion Questions

1. Why do standard preprocessing techniques need to be further customized to a particular corpus?
2. Did you find evidence for the idea that Democrats and Republicans have different sentiments in their tweets? What about Senators and Representatives?

3. Why is validating your exploratory and unsupervised learning approaches with a supervised learning algorithm valuable?

4. Did text only, non-text only, or text and non-text features together perform the best? What is the intuition behind combining text and non-text features in a supervised learning algorithm?

9 Large Language Model and Text Generation

Use GPT-2 or any other Large Language Model (LLM) with the Hugging Face library to generate few words using input text for your choice. In the lab, we generated text using input text: "I study sociology at University of California." Use suitable input text and generate very short text to describe this project. We understand that these models are not perfect and we are not expecting the generated text to be perfectly coherent or sensible.

If time permits, you are encouraged to tune hyperparameters but are not explicitly required to do so.

You can use code from lab as a starting point for this task but be sure to submit a link on bCourse.