Aarya Kagalwala, Nico Rapallo, and Timor Osman
Professor Tucker
Natural Language Processing
May 13th, 2024
GitHub Repo Link: https://github.com/aa-kag/NLPolitical-Project.git

Natural Language Processing Application to Political Speech

Abstract

The primary research question of this project is whether a natural language processing approach could effectively categorize members of congress and other prominent figures based on political affiliation. Text of congressional press releases from the past month were used as training and validation data for a TF-IDF model and RNN model. The data was web scraped using BeautifulSoup from the Propublica website where politicians' press releases are released on a daily basis. 1000 press releases and their respective politicians, political affiliation (Democrat or Republican), and the title of the press release. After gathering the data some work had to be done to clean it up for the model. Upon review of the raw data that was scraped there was a need to remove nulls, clean up extraneous html information, and some press releases were not useful due to length. After cleaning the data, there were labels placed on select representatives who were part of 4 different caucuses: Progressive, New Democrat, Main Street, and Freedom. These select representatives were used in a subset dataframe for further analysis. On the full data frame a TF-IDF was run with the scores outputting around 80-87% accuracy. Additionally, to begin attempting to classify documents by party, an RNN model was run. The model consistently ran between 82 to 89% accuracy, indicating a strong model. A similar methodology was used on the Caucus specific data, where the TF-IDF accuracy was 70 to 80% and the RNN accuracy was around 56%. The party and caucus models were also tested on Biden and Trump press releases. The model tended to classify the press releases accurately, based on the word use within the press release texts.

Introduction

Politicians have to self-classify themselves by political party and caucus. Outside observers also categorized members of congress based on how they vote. Examples of such voting-based categorizations include Poole and Rosenthal's dw-nominate and 538's "The 8 Types Of Democrats And Republicans In The House." Our aim is to provide an NLP based metric by which a politician's statements can be judged. This has the additional benefit of being applicable to those who are not members of congress. We hope to also be able to use the model to measure how a politician's statements differ over time, perhaps throughout a campaign season, or on different issues.

Literature Review

There are existing studies that have utilized a variety of probabilistic models to observe trends in political rhetoric via the use of text-data. One such study completed by the Paul G. Allen School of Computer Science and Engineering sought to derive the ideological position of candidates in the 2008 and 2012 presidential election via text data from political speeches. They were able to establish and observe accurate trends that occurred during those election cycles by applying a hidden Markov Model on a corpus that was labeled via ideological speech cues. They tested their model by establishing pre-registered hypotheses of a candidate's political position and comparing those hypotheses with the outputs of the model.

Other studies regarding political positioning prediction (based on text data) have established certain behaviors of NLP models of political textual analysis. These include evaluating the effectiveness of predictive modeling on short vs long chains of text with longer text chains being observed to be much more accurately predicted . To combat issues with predicting shorter text chains one specific study found that feeding a model aggregated short political statements to predict increased its accuracy on predicting a politician's political party affiliation

Methodology and Dataset

We considered multiple sources for acquiring significant amounts of text data associated with politicians. The congressional record, a transcript of all proceedings and debates in congress, was the first place we considered for data collection. However, isolating the specific debates and speeches from the procedural language proved to be difficult. Twitter was another option for collecting public statements from political figures. The API on Twitter made this difficult, as well as fear of bias from certain members of congress disproportionately posting on twitter compared to others. The data we ultimately decided to use to construct our dataset, were press releases from every member of the House of Representatives and Senate. We collected this through a ProPublica database, which was updated with links to every congressperson’s recent press releases. From the ProPublica database, we scraped the links to the press releases, as well as the congressperson that published them and their political party. We then scraped the text of each press release from the linked websites and compiled each press release to its corresponding representative and political party. Once the dataset was constructed, we also labeled select representatives based on 4 caucuses, the House Progressive Caucus, New Democrat Caucus, Main Street Caucus, and the Freedom Caucus. We chose these four caucuses because they were the largest caucuses in the house which had the least amount of overlap between members (we dropped overlapping members from the caucus analysis). These caucuses also represented a significant amount of ideological diversity. We subsetted the representatives who were members of these caucuses into another dataset which would be used for further analysis.

Once the datasets were created they needed to be cleaned. To begin, a standard preprocessing function was run on the press releases in the datasets. The text was put into lowercase and punctuation was removed. The data was tokenized by document because it provided more context, gave more holistic meaning, and the texts were generally not long enough to warrant tokenization of sentences. When performing exploratory analysis, and reading through the data

sets, it was found that some of the rows contained “failed to scrape” so those rows were dropped from the data set. There were also 10 independent party press releases, but that representative, Angus King, caucasus with the democratic party so we relabeled him as a democrat.

After our data set was constructed, we conducted some exploratory data analysis and created visualizations which helped understand our data’s distribution as well as any mistakes with the dataset.

We began analysis of each of our constructed databases by running a TF-IDF on our full dataset (all of the members of congress with a press release) and our caucus dataset (members of our whole dataset who are a part of one of the four selected caucuses for our analysis). We used the TF-IDF to establish a baseline performance to compare our future models to. We then created an RNN model for each of these datasets, experimenting with a variety of hyperparameters adjusting based on performance.

Results

The first TF-IDF model that we ran was run on all of the congressmen that we web scraped who had press-releases. The TF-IDF score of this dataset was 87% accurate at predicting the party affiliation of all of the congressmen with relevant press releases. We then ran an RNN model to also predict for party affiliation on the same dataset and it was able to predict with 86% accuracy on the validation data. The 2 models had relatively similar accuracy when predicting party affiliation on the full list of members of congress with usable press releases.

Next we ran 2 models on the dataset we created with the members of congress who are members of one of the 4 caucasus that we chose for our research (Main Street Republican, Freedom, Progressive, and New Democrat). For the caucus data our models predicted based on the caucus affiliation label of the members of congress in that dataset. The TF-IDF for this data was able to predict caucus affiliation with 82% accuracy. Next the RNN model that we created was able to correctly predict caucus affiliation with up to 56% accuracy. Here 56% represents a prediction accuracy which is significantly higher than random chance because there are 4 possible caucus labels that could be predicted meaning random chance would be 25%.

To test our model we used the recent statements of a few prominent politicians, Joe Biden, Donald Trump, and Glen Youngkin. We used the ten most recent press releases for each, and in the case of Donald Trump ten most recent Truth Social posts, as his campaign does not seem to publish press releases for the public. For the most part, Joe Biden’s statements scored firmly Democratic throughout his statements, and was most associated with the Progressive Caucus, averaging 39% of the votes for the progressive caucus. Trump on the other hand strongly averages out as Republican (65%) and Freedom Caucus (54%). This is probably due to Trump’s Truth Social Post being more political, as well as his generally more inflamed rhetoric.

Youngkin on the other hand, happened to make almost entirely apolitical press releases in his previous ten. His texts were found to be not strongly democrat or republican averaging out at 52% democratic. This result makes sense as none of his press releases were expressly political. The caucus model predicted 31% for progressive, which while uncomfortably high, was still significantly lower than Biden's, and the other three caucus options remained just below 25, a sign that the data did not have strong political bias.

Discussion

Our model was most successful predicting very political and partisan speech. This is most likely due to the fact that the biggest differentiation between democrats and republicans is how partisan their speech is. More specifically with caucus prediction, members in the Freedom Caucus, which tend to have more partisan rhetoric are able to be more easily predicted. This is demonstrated in Trump's predicted caucus breakdown which was predicted to be overwhelmingly republican.

Our model has a few places where it falls short. The model does relatively well at predicting the political positions of the author, however it naturally struggles with more apolitical text, such as that which was released by Youngkin. The caucus level model also has a clear bias toward labeling as progressive. This is likely due to having disproportionately more training data from the progressive caucus than the other caucuses. Also, due to the model only being trained on the previous month (most recent 1000 press releases), it might lack the ability to decipher political positions which were written outside this time frame.

Our models and results for the most part agree with the other research that has been done in the field of political text analysis through natural language processing. Specifically, as with many other studies, we were able to predict political parties highly accurately. Additionally the challenges that we faced with the models we constructed (especially political nuance recognition) have also been highlighted by the literature surrounding political textual analysis.

Our research does however provide insight into a more specific domain of the political sphere in that we were able to somewhat accurately predict sub-categories beyond simply political parties (caucuses). In most research in the field that was reviewed, subcategories beyond simply political party affiliation or ideology were not explored deeply or were unable to be accurately explored through the use of textual analysis.

We see a few avenues for potential future research. The caucus model could be expanded by first clustering members of congress based on their voting history, and then using those clusters for the target of our RNN instead of caucus. Another potential avenue for more specific research, is filtering the training data by certain keywords which relate to topical issues. For example, only train a model on press releases which include the word 'Israel' or the phrase Inflation Reduction Act.' Through this method, we could measure attitudes on specific political issues.

Another research avenue that could be pursued using similar techniques, is measuring partisanship of a candidate's speech throughout a campaign season. To achieve this the training data would need to encompass a larger time frame, so that it could accurately predict over the duration of a campaign.

Conclusion

Our RNN model, based on congressional press releases, was able to successfully predict the party at a very high accuracy, and caucus at a relatively high accuracy. It performed best with text that was clearly political and partisan. The model struggled to predict political nuance (this is intuitive) and this is specifically observed in our RNN's poor prediction of the new democrat caucus.

References

"Measuring Ideological Proportions in Political Speeches." *Measuring Ideological Proportions in Political Speeches | Paul G. Allen School of Computer Science & Engineering*, www.cs.washington.edu/publications/measuring-ideological-proportions-political-speeches. Accessed 13 May 2024.

Measuring Political Positions from Legislative Speech, benjaminlauderdale.net/files/papers/2016LauderdaleHerzogPA.pdf. Accessed 14 May 2024.

Predicting Political Party Affiliation from Text - Sebastian Schelter, ssc.io/pdf/poltext.pdf. Accessed 14 May 2024.

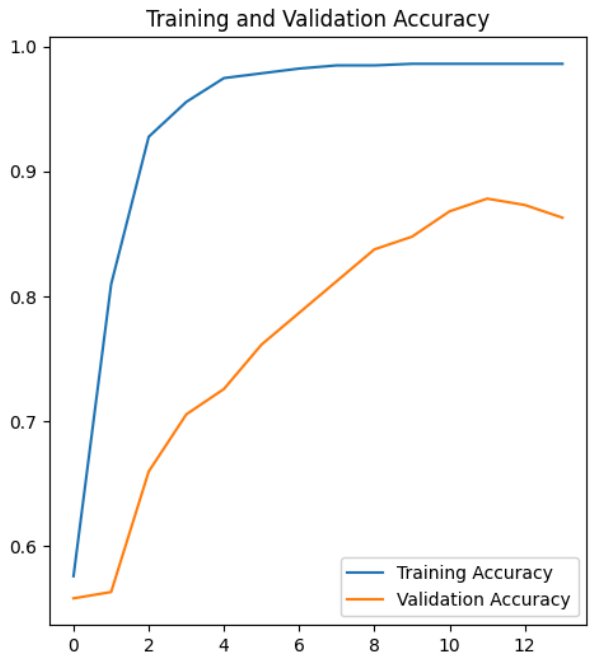
Tia Yang, Cooper Burton. "The 8 Types of Democrats and Republicans in the House." *FiveThirtyEight*, 1 May 2024, projects.fivethirtyeight.com/types-democrats-republicans-house-2024/.

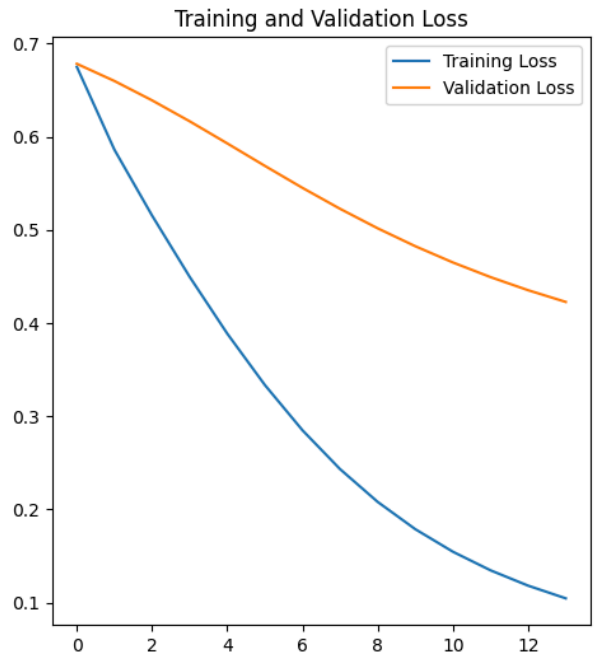
"UCLA Presents Voteview.Com Beta." *Voteview*, voteview.com/about. Accessed 13 May 2024.

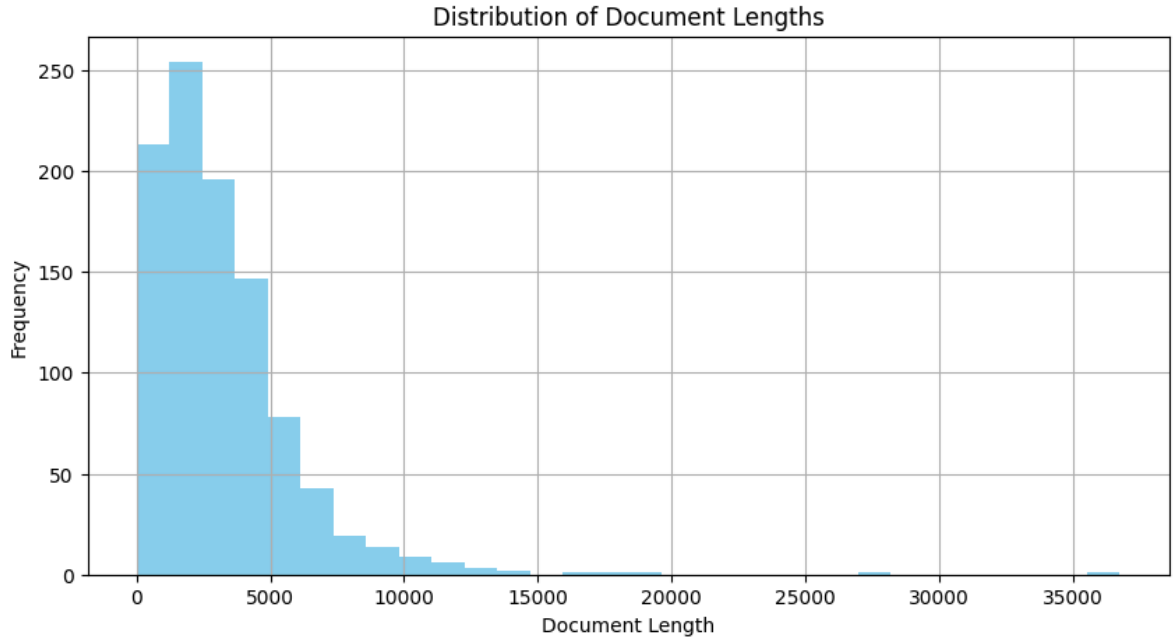
Willis, Derek, et al. "Represent." *ProPublica*, 12 Aug. 2015, projects.propublica.org/represent/statements?page=1.

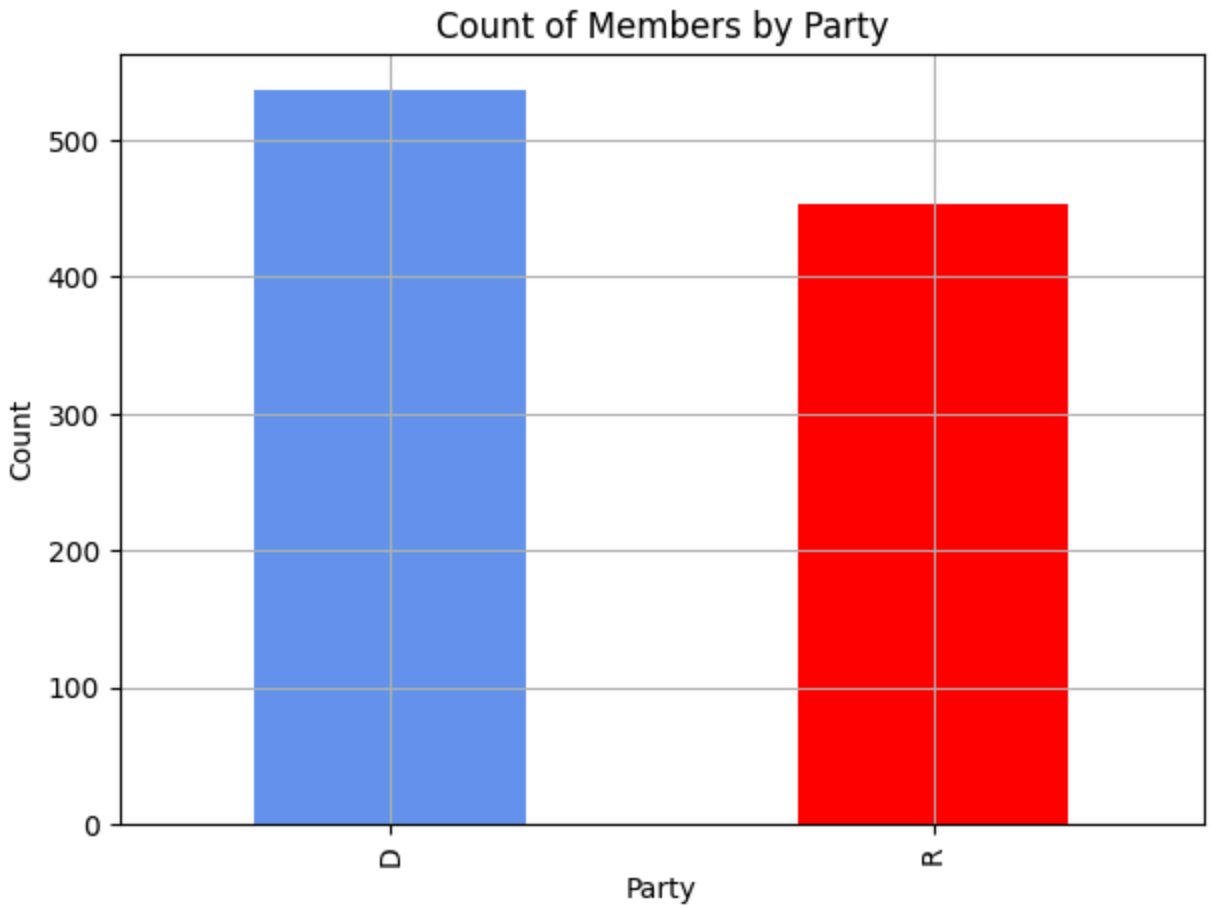
Visuals referring to our findings and EDA:

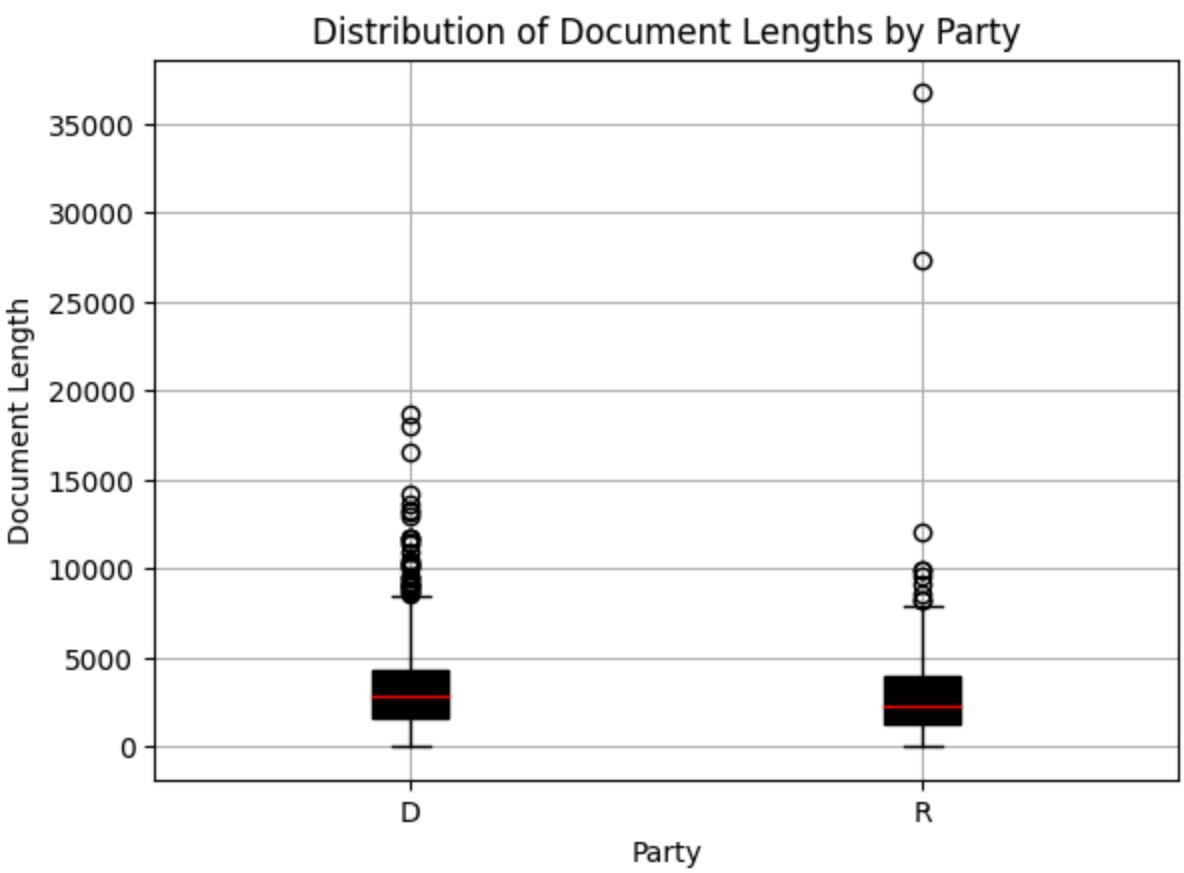
RNN Performance based on the full data:

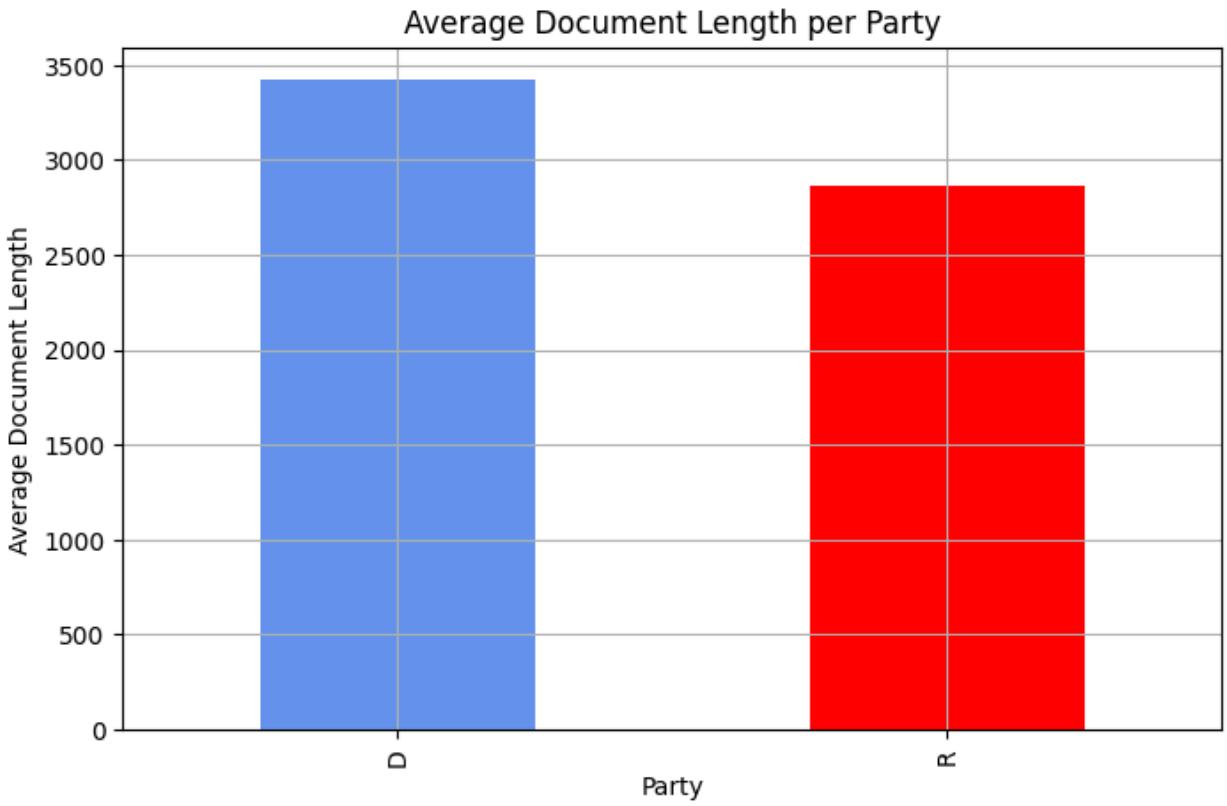

Training and Validation Accuracy
Training Accuracy
Validation Accuracy

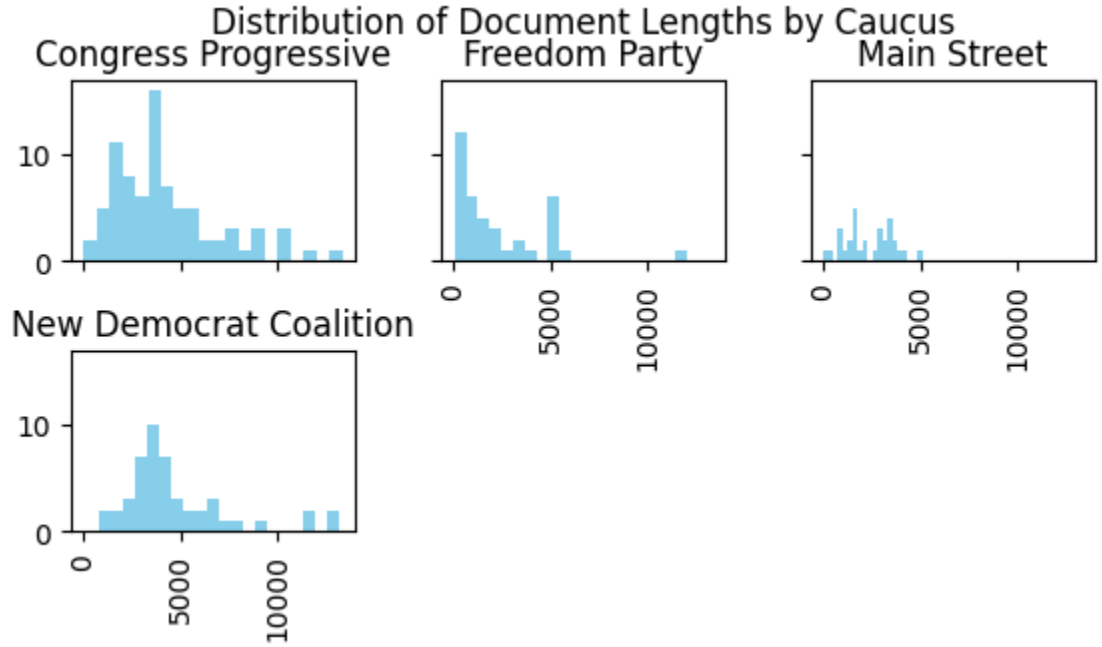

Training and Validation Loss
Training Loss
Validation Loss

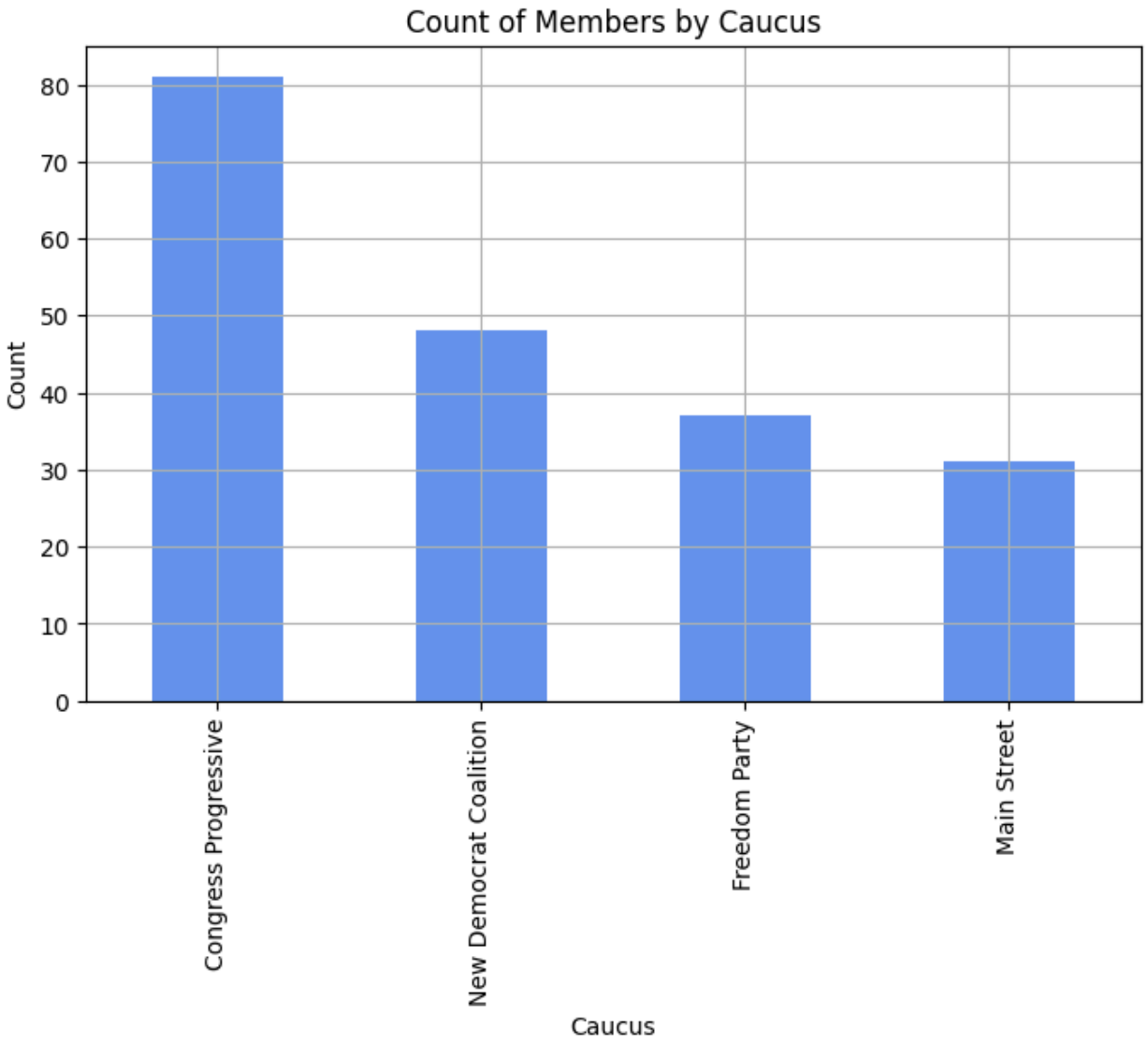

Distribution of Document Lengths
Frequency
Document Length

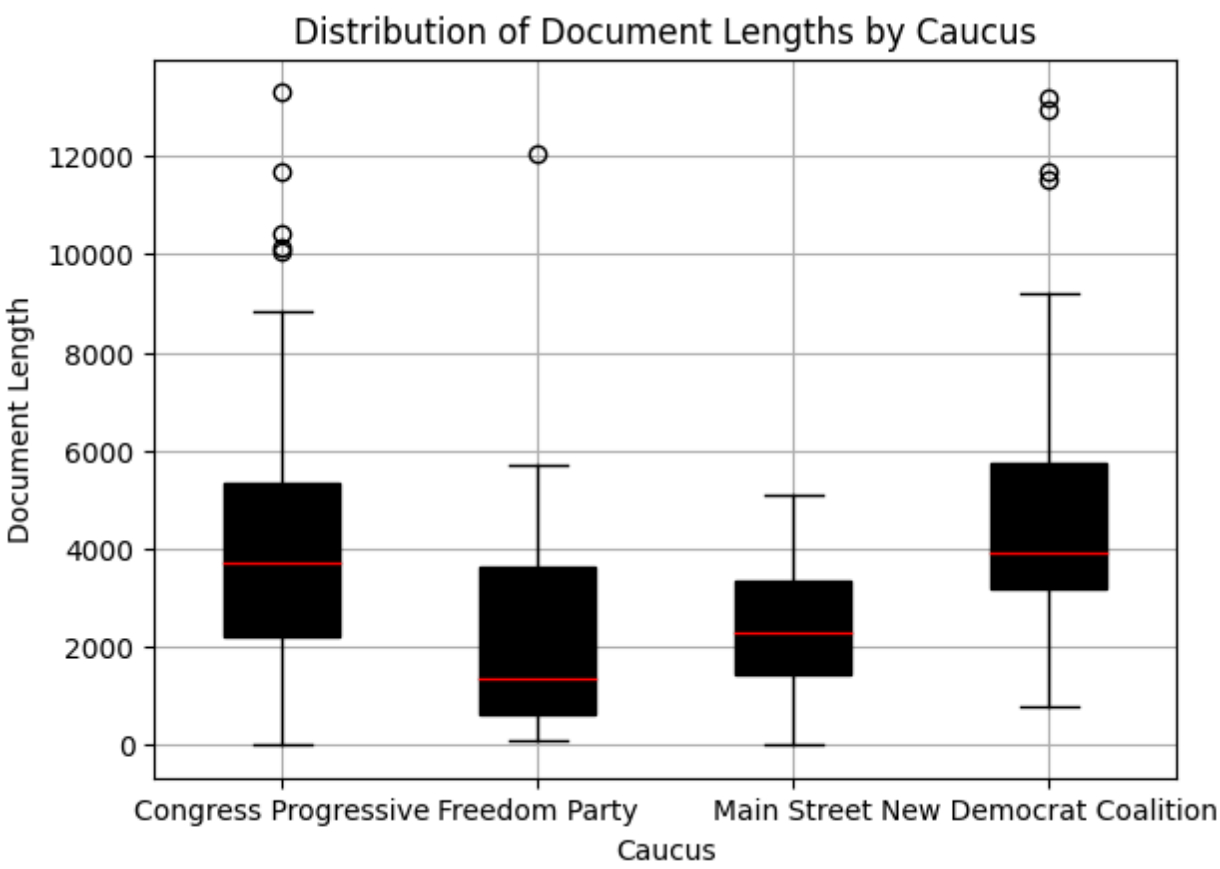

Count of Members by Party
500
400
300
200
100
0
Count
D
R
Party


Distribution of Document Lengths by Party
Document Length
35000
30000
25000
20000
15000
10000
5000
0
D
R
Party


Average Document Length per Party
Average Document Length
3500
3000
2500
2000
1500
1000
500
0
D
R
Party


Distribution of Document Lengths by Caucus
Congress Progressive
Freedom Party
Main Street
New Democrat Coalition
10
0
0
5000
10000


Count of Members by Caucus
Count
0
10
20
30
40
50
60
70
80
Congress Progressive
New Democrat Coalition
Freedom Party
Main Street
Caucus


Distribution of Document Lengths by Caucus
Document Length
0
2000
4000
6000
8000
10000
12000
Congress Progressive
Freedom Party
Main Street New Democrat Coalition
Caucus