Sasha Matveev

IS 417

Dr. Underwood

5 December 2021

Divided We Fall: Using Machine Learning to Analyze U.S. Senate Political Polarization

**Introduction**

The growing political polarization, or mean difference between party ideology has seemingly seeped its way into every household and every aspect of daily life in the United States. According to the latest Pew Research Poll, 59% of Americans say it is stressful to have a conversation about politics with a person they ideologically disagree with. This reflects nearly a ten percent increase from the same poll taken in 2019 (Green). The lack of bi-partisan bills and legislation passed in recent years has also decreased as Congress’ approval ratings have been hitting historic lows in the single digits since the mid-2010s (Gallup). Although social media and the speed of information technology may have exacerbated the trend and thrust it firmly into the mainstream, increasing political polarization has been tracked for over 20 years. The ideological median of both the Democratic and the Republican parties has been shifting towards their respective extremes. Since 1994, the median ideology has shifted over 24% away from neutral for Democrats and over 28% for Republicans (Pew Research Center).

Starting in the 1980s, political scientists Keith T. Poole and Howard Rosenthall began quantifying legislative roll-call voting patterns in empirical terms with a Nokken-Poole two-dimensional score. Simply put, the score is a number that reflects the strength of a politician’s association with a liberal or conservative ideology. The goal of this study is to utilize machine learning to examine whether the speeches given by the U.S. Senators reflect their ideology and whether or not the speeches they give are reflective of the increasing polarization over time as determined by the Nokken-Poole scores. Based on the variation in social contexts for senators and politicians at large, the second dimension of the Nokken-Poole score will likely be a less accurate predictor of party affiliation and partisanship. The economic ideology measured by the first dimension Nokken-Poole score should provide a more consistent representation of partisanship as bills targeting the economy demand party loyalty to a higher degree.

**Data**

The data for this project came from Harvard’s dataverse and Voteview, both of which were relatively easy to obtain and came fairly clean. Harvard’s dataverse contained csv files of vectorized speeches by U.S. Senators from the 104th-113th Congress while also containing metadata about the individual senators such as their party affiliation, birth year, home state, sex, along with many other categories. The vectorized words picked out from individual speeches are truncated to eliminate the tense of the word for the sake of clarity and simplicity. Unfortunately, the dataset was published in 2016 and does not contain any data from the latest administration.

The data from Voteview contained the empirical measurements of party affiliation and general polarization based on voting records over the same Congress sessions; these were quantified as Nokken-Poole nominate scores which range from negative one to positive one. Negative one represents extreme conservative views while positive one represents extreme liberal views. The scores have two dimensions: the first dimension scores the politician based on their economic outlook: liberal or conservative, while the second dimension scores the politician based on their cultural and social outlook. Neither of the data sources provided a concrete count as far as the number of speeches and their length; as a result, the data may be skewed somewhat toward the “loud minority” as opposed to the “silent majority”. Due to the fact that nearly 1000 congressmen and congresswomen are contained within the vectorized speeches dataset, there are enough data points across the ten U.S. Congress sessions to obtain a statistically significant result.

A significant challenge in working with the datasets was combining the two while simultaneously extracting the valuable information contained within. The Voteview and Harvard dataverse data came in csv format and the two files were indexed and organized differently which compounded the difficulty of data processing and preparation. To solve the problem, a mediator file was created that contained the two variables used to index each dataset. The two datasets were then merged on the index of the mediator file to create a Python Pandas data frame object that now contained both the Nokken-Poole scores and the index associated with each senator in their speech file. The use of a simple join, merge, or concat function did not produce the desired result as the names of each senator were in different formats in the two data sets.

**Methods**

For the sake of simplicity, members of defunct parties or political independents were grouped with the party they generally caucus or otherwise associate with. Liberals generally identifying with the Democratic party were given a score of one while conservatives generally identifying with the Republican party were given a score of zero upon grouping. Due to time constraints, only the 108th Congress, which was in session from 2003-2005 was analyzed. The revelations from that congressional session should offer an insightful glimpse into a time of similar political turmoil and serve as a parallel to the current, tense political climate.

Unlike four algorithms Andrew Peterson and Arthur Spirling used in their analysis of political speeches from Westminster, this study focuses on logistic regression while utilizing the method predict\_proba to generate a probability that the words spoken by a particular congressman are indicative of their liberal (Democratic) or conservative (Republican) association. The method predict\_proba returns two values: the first is the probability that the text is associated with conservative ideology (closer to 0), the second is the probability that the text is associated with liberal ideology (closer to 1). Since the original sorting method utilized 1 as the ceiling, the second value was chosen as a representative indicator of polarization from speeches. The 108th congress contained speeches from 99 members so the dataset was randomly split with ten members in the test dataset and the other 89 in the training dataset.

In sorting between false positives, false negatives, true positives, and true negatives, a threshold of 0.5 was used since the original sorting method set 1.0 as the extreme liberal value and 0.0 as the extreme conservative value. The actual value of 0.5 was included in the liberal spectrum although not a single candidate was exactly at that value based on the text analysis done by the logistic regression model. To ensure the validity of the results, the cross validation method cross\_val\_predict was implemented with a C parameter of 0.001 and the predict\_proba method once again. The cross validation method returned an array of two values, with the first value being the probability of 0 and the second value being the probability of 1. As before, the second value (probability of 1) was taken. An F1 score was calculated as shown in Figure 1.

Measuring Accuracy of Logistic Regression

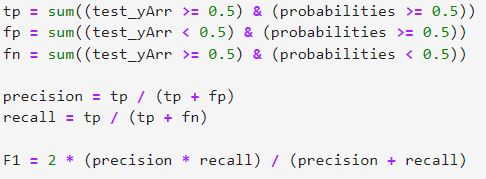


Figure 1: Calculation of the F1 score.

Following the successful cross validation of the results, an F1 score of 0.833 was obtained which suggests a rather high degree of accuracy in the logistic regression model.

**Results**

The last step in determining whether the speeches of senators were a reliable way of determining their political ideology and partisanship as dictated by their voting patterns was to find the statistical significance of the obtained results. Figures 2 and 3 show two simple scatter plots of the variables logistic regression probability of 1 vs Nokken-Poole dimension 1, and logistic regression probability of 1 vs Nokken-Poole dimension 2 were constructed.

Nokken-Poole First Dimension Score vs Logistic Regression Model Probability

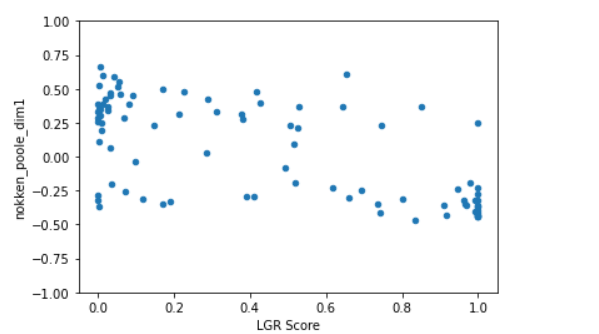


Figure 2: Scatter plot of Nokken-Poole first dimension score vs the model predicted probability.

Nokken-Poole Second Dimension Score vs Logistic Regression Model Probability

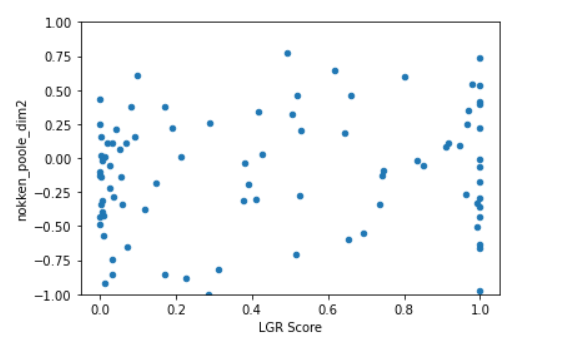


Figure 3: Scatter plot of Nokken-Poole second dimension score vs the model predicted probability.

Although the two variables were scaled differently, a relationship between them can still be seen in Figure 2. To empirically verify the relationship between the three variables, Python’s pearsonr correlation method was used. The method outputs the r-value which indicates the strength of the linear relationship between the variables, as well as the p-value which is necessary to verify the statistical significance of the results obtained in the study. For both methods, n was kept at 90 senators. As seen in Figure 4, the p-value of the logistic regression probability and the first dimension Nokken-Poole score was a staggering 7.416x10^-12 which is effectively zero and certainly less than 0.05, making it incredibly statistically significant.

Examining the Strength of Linear Relationships and Statistical Significance

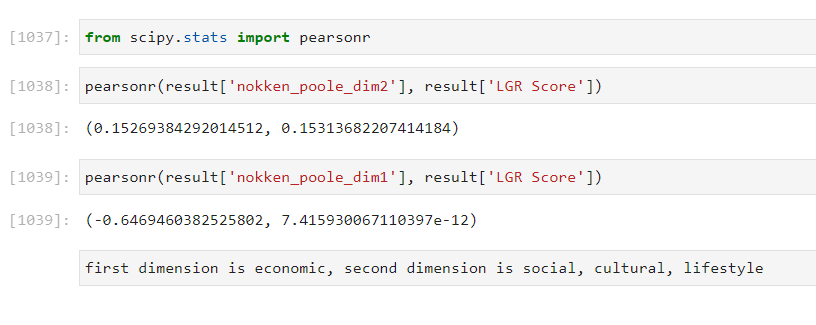


Figure 4: Calculated r-values followed by p-values for second and first dimension Nokken-Poole scores, respectively.

The r-value is -0.646 indicating a moderately strong linear relationship between the first dimension score and the model predicted probability. The p-value between the second dimension Nokken-Poole score was 0.153 which is greater than 0.05, making the result statistically insignificant. The r-value for the second dimension is 0.153 indicating that there is a very weak linear relationship between the second dimension score and the model predicted probability.

Generally speaking, the first dimension Nokken-Poole score is considered to be the most accurate and precise measure of a politician’s partisanship as a significant portion of crucial votes on economic policy rely heavily on party loyalty. Party members are generally more united by economic ideology rather than cultural or social ideology and values: a trend that is consistent with the results of this study. The trend can also be observed within the current 117th U.S. Congress as Republicans are culturally divided between the “traditional” or “old-school” politicians like Liz Cheney and Mitt Romeny and the hyper-partisan “new-school” politicians like Marjorie Taylor Greene and Lauren Boebert. While they were all united in opposition to economic bills by Democrats, their social values differ significantly as Liz Cheney is on the committee investigating the Capitol Riot while the new-school politicians generally allied with former President Trump oppose the investigation in some capacity.

**Conclusion**

The cross-validated F1 score of 0.833 for the 108th Congress along with a statistically significant p-value of 7.416x10^-12 for the relationship between the first dimension Nokken-Poole score and the logistic regression model’s probability prove the initial hypothesis posed at the start of the study. The model does indeed predict the political ideology, and by extension, the partisanship of senators with a reliable degree of statistical significance. As discussed earlier, economy-related bills often carry more weight in terms of their impact on the nation so it would make sense that party leaders would whip up votes to push their agenda and enforce party loyalty. In contrast, bills that deal with social causes generally see more variation in party loyalty as no two senators or politicians likely see eye-to-eye on every social issue. There are also likely Democrats that are economically liberal but socially conservative, and vice-versa. The model used in this study does not account for this “crisscrossing” of ideological lines as each dimension of the Nokken-Poole score is calculated independently of each other and of other sessions of Congress. Thus, it is understandable that the second dimension Nokken-Poole score would be worse than the first dimension score in predicting party loyalty.

Another factor to take into account is the context surrounding the speeches that were vectorized to conduct text analysis. It would make sense that economic policy would invite more debate and more speeches by senators arguing their viewpoint in comparison to social issues which typically hold secondary priority. As a result, the data used to analyze the speeches may be somewhat biased towards accuracy with the first dimension Nokken-Poole score due to the content of the speeches. Without obtaining the actual speeches used in the dataset, it is impossible to tell how much of a factor this was in the study.

Due to time constraints and overall scope of this study, only the speeches from the 108th session of the U.S. Congress were analyzed. The greatest limitation to this study is its lack of contextualization over the time period originally proposed. Even though the 108th Congress does offer a valid parallel to today’s circumstances, with trouble at home and abroad, it is certainly no substitute for a thorough analysis over time. It is reasonable to suspect that the logistic regression model that worked so well during this study would also produce a statistically significant result if it were applied to other sessions of Congress. The main point of interest over time, aside from tracking the degree of polarization in speeches, would be to track the statistical significance of the second dimension Nokken-Poole score and see if it improves or worsens as the two parties become more polarized over time. One would assume that as the two parties grow further apart, social issues would begin to factor in behind the wedge being driven between Democrats and Republicans.

Future studies should explore the validity of the model in more recent datasets (114th Congress and on) to see if the model is still accurate and if the model’s accuracy has improved with the increasing polarization. Furthermore, future studies should explore possible applications of the model in the U.S. House of Representatives and potentially in other democracies abroad. It is unclear whether or not the U.S. House of Representatives would offer the same degree of valid data as there are nearly four times the amount of members within the House and it is unrealistic to think that every member gives speeches on the House floor. The data could be skewed towards the vocal majority to a greater degree than is possible in the U.S. Senate as a result of the sheer size of the House of Representatives.

An additional area of analysis could take inspiration from the study done by Andrew Peterson and Arthur Spirling titled “Classification Accuracy as a Substantive Quantity of Interest: MeasuringPolarization in Westminster Systems” and utilize multiple analytical machine learning algorithms, picking the one with the greatest accuracy for the legislative session. This would likely increase the statistical significance of the results as the logistic regression method of analyzing text is far from perfect and does have its limitations. Lastly, it may be interesting to take into consideration the social media posts by various politicians to see how they may support or contradict the positions taken during official speeches. Not every politician has a social media presence so the problem of inaccurate data due to a vocal minority is very likely to arise yet again.

Works Cited

“Congress and the Public.” *Gallup*,2021,https://news.gallup.com/poll/1600/congress-public.aspx.

Green, Ted Van. “Republicans and Democrats Alike Say It’s Stressful to Talk Politics with People Who Disagree.” *Pew Research Center*, 23 Nov. 2021, https://www.pewresearch.org/fact-tank/2021/11/23/republicans-and-democrats-alike-say-its-stressful-to-talk-politics-with-people-who-disagree/.

Lauderdale, Benjamin E.; Herzog, Alexander. *Replication Data for: Measuring Political Positions from Legislative Speech*. (Data file version 1). Harvard Dataverse, 2016. Web. 5 Dec 2021. <<https://doi.org/10.7910/DVN/RQMIV3>>

Lewis, Jeffrey B, et al. *Voteview: Congressional Roll-Call Votes Database.* (Data file). Voteview, 2021. Web. 5 Dec 2021. <https://voteview.com/data.>

Peterson, Andrew, and Arthur Spirling. “Classification Accuracy as a Substantive Quantity of Interest: Measuring Polarization in Westminster Systems.” *Political Analysis*, vol. 26, no. 1, 2018, pp. 120–128., doi:10.1017/pan.2017.39.

“Political Polarization in the American Public.” *Pew Research Center*, 12 June 2014, https://www.pewresearch.org/politics/2014/06/12/political-polarization-in-the-american-public/.