GEORGETOWN UNIVERSITY
MASSIVE DATA INSTITUTE
McCourt School of Public Policy

MDI Scholars Spring 2020
Project Summary

Assessing the Ideology of Democratic Congressional Candidates
Faculty Lead: Michael A. Bailey
MDI Scholars: Andy Green, Neha Tiwari

Abstract

This project aims to assess the ideology of Democratic congressional candidates by analyzing the language used on the candidates' campaign websites. We start out by using web scraping techniques to collect text data from the campaign websites of both non-incumbent candidates and incumbent House members. After collecting the text data, we utilize cosine similarity and word collocation analysis to score each candidate or member on an ideological scale. While the project is still a work in progress and final results are not yet available, the preliminary results are promising. If the methodology can be further refined and validated, we hope that this type of analysis can serve as a supplemental measure to existing tools for assessing political ideology.

Introduction

This project aims to assess the ideology of Democratic candidates running in the 2020 elections for the United States House of Representatives. Traditional measures of political ideology, such as DW-NOMINATE scores, utilize congressional voting records to score ideology,[1] and thus cannot be applied to non-incumbent candidates. Additionally, even for incumbent representatives, DW-NOMINATE scores can be thrown off by "protest" votes, producing counter-intuitive results for members considered to be more ideologically extreme.[2] As a result, developing new methods for evaluating political ideology can be beneficial for evaluating non-incumbent candidates, as well as for providing supplemental measures to existing techniques for assessing ideology amongst incumbents.

---

[1] Voteview. "About the Project." *Voteview*, voteview.com/about.
[2] Lewis, Jeff. "Why Is Alexandria Ocasio-Cortez Estimated to Be a Moderate by NOMINATE?" *Voteview*, 5 Aug. 2019, voteview.com/articles/ocasio_cortez.

Methodology

Data Sources and Collection

In order to assess the ideology of congressional candidates, we utilize web scraping techniques to gather text data from each candidate's campaign website. While there are significant differences in how each candidate structures their website and what types of content are included, the vast majority of the websites have a few key sections in common. Specifically, we focus on the following portions of a candidate's campaign website:

- "About" page, where candidates typically introduce themselves, their background, experience, and values
- "Issues" page(s), where candidates typically discuss the issues they consider to be most important, how they would address those issues, and their qualifications or achievements relevant to those issues
- "Home" page, in the case that it contains substantive information about their background or stances on issues that cannot also be found on the "About" or "Issues" pages

Over the course of the project, we scraped the websites of 111 total individuals, split between 91 non-incumbent candidates and 20 incumbent members. The non-incumbent candidates were chosen by focusing on two distinct categories of seats/races:

- Seats currently held by a Democratic incumbent who is not seeking reelection in 2020
    - Theoretically, these races should produce some of the most robust and competitive Democratic primary elections, as the incumbent isn't running, and many are considered "safe" Democratic seats.
- Seats currently held by an incumbent of either party who is not seeking reelection in 2020, where the general election is also likely to be competitive[3]
    - Theoretically, these races should also produce a relatively robust and competitive primary election, as the incumbent isn't running, and the seat is considered to be attainable for the party in the general election.

Between these two groups, we identified 17 races to focus on. After obtaining a list of all candidates registered to run in those races,[4] we collected data from all candidates who had a website, resulting in a total of 91 candidates.

The group of 20 incumbent members can be split into two categories:

---

[3] Cook Political Report. "2020 House Race Ratings." *The Cook Political Report*, 24 Apr. 2020, cookpolitical.com/ratings/house-race-ratings.

[4] Wikipedia. "2020 United States House of Representatives Elections." *Wikipedia*, 14 May 2020, en.wikipedia.org/wiki/2020_United_States_House_of_Representatives_elections.

- 10 members considered to be among the most liberal members of the Democratic caucus
- 10 members considered to be among the most moderate members of the Democratic caucus

In selecting the members for the liberal and moderate groups, we relied primarily on DW-NOMINATE scores, selecting from members clustered at either ideological pole of the Democratic Party. However, due to the concerns with the DW-NOMINATE scale discussed earlier, we also relied upon other factors like caucus membership (e.g. Congressional Progressive Caucus[5] or the Blue Dog Coalition[6]), as well as conventional wisdom and informal groups (e.g. the "squad"[7] or the "badasses"[8]).

Pre-Processing

After collecting data on the candidates and members, the next important step was to prepare the data for analytical purposes. In order to effectively carry out the analyses described below, there are a number of manipulations that need to be applied to the text data first. These manipulations include actions like removing all punctuation, converting all letters to lowercase, and removing common "stop words" (e.g. "the", "it", "and", etc.). Research from Denny and Spirling shows that the specific pre-processing actions that are taken can have a significant effect on the ultimate analytical results.[9] In order to ensure we were using the optimal pre-processing steps, we used the authors' "preText" package, which helps provide guidance on the ramifications of using various pre-processing techniques. The analysis indicated that we should proceed with the three methods mentioned above (removing punctuation, lowercasing, and stop word removal), as well as stemming, which reduces all words to their "stem" form (e.g. "run", "running", and "runs" would all be treated as the same word).[10] As a result, we implemented these four pre-processing methods on the text data before proceeding to the analyses discussed below.

Analysis

After completing the text pre-processing, we engaged in two primary streams of analysis – cosine similarity and word collocation analysis.

---

[5] Congressional Progressive Caucus. "Caucus Members." *Congressional Progressive Caucus*, cpc-grijalva.house.gov/caucus-members/.
[6] Blue Dog Coalition. "Members." *Blue Dog Coalition*, bluedogcaucus-costa.house.gov/members.
[7] Sullivan, Kate. "Here Are the 4 Congresswomen Known as 'The Squad' Targeted by Trump's Racist Tweets." *CNN*, 16 July 2019, www.cnn.com/2019/07/15/politics/who-are-the-squad/index.html.
[8] Bash, Dana, and Bridget Nolan. "These Five Freshman Congresswomen Changed History by Becoming Unlikely Leaders on Impeachment." *CNN*, 28 Sept. 2019, www.cnn.com/2019/09/28/politics/badass-women-impeachment-democrats-oped/index.html.
[9] Denny, Matthew J, and Arthur Spirling. "Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It." *NYU*, 27 Sept. 2017, www.nyu.edu/projects/spirling/documents/preprocessing.pdf.
[10] Ibid.

In the context of text data, cosine similarity can be used as a mathematical means of determining how similar two pieces of text are to one another,[11] with values ranging from 0 (nothing in common), to 1 (exactly the same). We created a matrix whereby the cosine similarity would be calculated between each candidate/member and every other candidate/member. To help visualize this matrix, a small subset of the total matrix is displayed in Figure 1 below.

Figure 1. Subset of the cosine similarity matrix

	grijalva_raul	caballero_jose	vazquez_joaquin	goldbeck_janessa
grijalva_raul	1	0.382	0.614	0.618
caballero_jose	0.382	1	0.432	0.399
vazquez_joaquin	0.614	0.432	1	0.548
goldbeck_janessa	0.618	0.399	0.548	1

For each candidate/member, the primary data points of interest are the candidate/member's cosine similarity with the 10 liberal members and the 10 moderate members. As such, we calculated the average of each candidate/member's score amongst each of these groups. In the case of the members, their cosine similarity score against their own text was excluded from the calculation, as this would artificially inflate their true score. Finally, we took the difference between each candidate/member's average scores amongst the liberal and moderate members, resulting in a single "net" score for each candidate. This score represents how "similar" that candidate/member is to those on either side of the ideological spectrum, based on the language found on their campaign websites.

Separate from the cosine similarity analysis, we also conducted word collocation analysis. Collocation analysis works by identifying words that appear together frequently in a piece of text.[12] In this analysis, we focused specifically on groupings of either two or three words, referred to as bigrams and trigrams, respectively. We identified a group of bigrams/trigrams that the liberal members were disproportionately likely to use (e.g. "Medicare for All" and "Green New Deal"), as well as a group of bigrams/trigrams that the moderate members were disproportionately likely to use (e.g. "get things done" and "both parties"). After sifting through the results of the collocations analysis, we identified 45 politically-charged bigrams/trigrams to focus on.

Additionally, we identified 21 unigrams (i.e. single words) that were used disproportionately frequently by the liberal and moderate members (e.g. "housing" and "women" for liberal members, and "tax" and "job" for moderate members). Combining

---

[11] Gupta, Sanket. "Overview of Text Similarity Metrics in Python." *Towards Data Science*, 15 May 2018, towardsdatascience.com/overview-of-text-similarity-metrics-3397c4601f50.

[12] Ruchirawat, Nicha. "Collocations - Identifying Phrases That Act like Single Words in Natural Language Processing." *Medium*, 16 Mar. 2018, medium.com/@nicharuch/collocations-identifying-phrases-that-act-like-individual-words-in-nlp-f58a93a2f84a.

the 45 bigrams/trigrams and the 21 unigrams, each of these n-grams was given a diagnostic score based on how many members from each group uses it (e.g. "Medicare for All" is used by 7/10 liberal members and 1/10 moderate members, leading to a score of -6). These diagnostic scores were then used to weight the relative importance of each n-gram in calculating a final score for each candidate based on the n-grams that they did or did not use. In other words, this score represents how "similar" each candidate/member is to those on either side of the ideological spectrum, based on specific politically-charged words and phrases found on their campaign websites.

Results

While this project is very much a work in progress, the preliminary results are generally promising. Starting with the cosine similarity analysis, the candidates that we subjectively perceived to be among the most liberal or moderate generally did score toward the expected poles. Perhaps a more useful validation check, though, is assessing the cosine similarity scores of the members. After all, if the liberal members don't generally score similarly to each other (and the same on the moderate side), there's good reason to question either the validity of the benchmark groups we've selected, or the validity of the methodology altogether.

Fortunately, the analysis performs reasonably well on this front. Six out of the 10 benchmark liberal members fall within the 15 most liberal scores (out of the total pool of 111 observations), while five out of the 10 benchmark moderate members fall within the 15 most moderate scores.

However, it's also clear that the analysis is far from perfect. While it seems to perform well in classifying some of the more liberal and moderate members, the model also produces some counter-intuitive results. Figure 2 below shows the 20 benchmark liberal and moderate members, sorted in order of their cosine similarity scores, from more moderate/right scores at the top, to more liberal/left scores at the bottom. If the model was working exactly as intended, it would produce a clean sorting with all moderate members at the top, and all liberal members at the bottom. In this case, we see some anomalies, with members like Ro Khanna – who was a co-chair for Bernie Sanders' presidential campaign[13] – scoring to the "right" of members like Abigail Spanberger, who is considered to be staunchly moderate.[14]

---

[13] Garofoli, Joe. "Bernie Sanders Enlists Ro Khanna for Presidential Campaign." *San Francisco Chronicle*, 21 Feb. 2019, www.sfchronicle.com/politics/article/Bernie-Sanders-enlists-Ro-Khanna-for-presidential-13635537.php.

[14] Portnoy, Jenna. "Rep. Abigail Spanberger: A Moderate Democrat Working to Survive in the AOC Era." *The Washington Post*, 28 May 2019, www.washingtonpost.com/local/virginia-politics/rep-abigail-spanberger-a-moderate-democrat-working-to-survive-in-the-aoc-era/2019/05/16/a2ff11e4-700c-11e9-8be0-ca575670e91c_story.html.

Figure 2. List of 20 benchmark liberal and moderate members, sorted by cosine similarity scores, from most moderate/right at the top, to most liberal/left at the bottom

Candidate	Group
gottheimer_josh	Incumbent_Moderate
sherrill_mikie	Incumbent_Moderate
luria_elaine	Incumbent_Moderate
mcadams_ben	Incumbent_Moderate
lamb_conor	Incumbent_Moderate
murphy_stephanie	Incumbent_Moderate
brindisi_anthony	Incumbent_Moderate
khanna_ro	Incumbent_Liberal
golden_jared	Incumbent_Moderate
fudge_marcia	Incumbent_Liberal
lipinski_dan	Incumbent_Moderate
spanberger_abigail	Incumbent_Moderate
jayapal_pramila	Incumbent_Liberal
watsoncoleman_bonnie	Incumbent_Liberal
bass_karen	Incumbent_Liberal
grijalva_raul	Incumbent_Liberal
ocasiocortez_alexandria	Incumbent_Liberal
omar_ilhan	Incumbent_Liberal
gomez_jimmy	Incumbent_Liberal
pressley_ayanna	Incumbent_Liberal

Another example of a counter-intuitive result produced by the model is with incumbent Dan Lipinski, and the (recently successful) primary challenger for that seat, Marie Newman. As someone who voted against the ACA and has been a longtime opponent of abortion rights, Dan Lipinski is viewed as one of the more conservative members of the Democratic caucus.[15] On the other hand, Marie Newman is viewed as solidly progressive, having earned the endorsements of liberal leaders like Alexandria Ocasio-Cortez.[16] However, the cosine similarity analysis actually places Marie Newman to the "right" of Dan Lipinski, a result that doesn't make a lot of intuitive sense.

It's also worth briefly discussing here another factor that could be affecting the cosine similarity results. There is a fairly large disparity in the length of the text data that we scraped from candidates' websites, ranging from 54 words on the short end to over 7,000 words on the long end (after removing stop words, as discussed in the

[15] Stolberg, Sheryl Gay. "Marie Newman Beats Dan Lipinski, Democratic Incumbent, in Illinois House Primary." *The New York Times*, 18 Mar. 2020, www.nytimes.com/2020/03/18/us/politics/marie-newman-dan-lipinski-illinois.html.
[16] Ibid.

Methodology section). This is potentially problematic for the cosine similarity analysis, as there appears to be a statistically significant correlation between the total number of words used and the candidate's "net" cosine similarity score, where having more words is associated with scoring more toward the liberal side. This may be related to the fact that the benchmark liberal members tend to use significantly more words than the moderate members on average. The liberal members use 2,758 words on average, while the moderate members use 1,654 words on average. (For reference, the average across the total sample is 1,240 words).

Shifting gears to the word collocation analysis, preliminary results from this workstream appear to improve upon some of the shortcomings of the cosine similarity analysis. Figure 3 below shows the same 20 benchmark liberal and moderate members as above, now sorted in order of their politically-charged n-gram scores, again with more moderate/right scores at the top, and more liberal/left scores at the bottom. Contrary to the cosine similarity results, the members now sort perfectly, with all 10 moderate members to the "right" of all 10 liberal members.

Figure 3. List of 20 benchmark liberal and moderate members, sorted by politically-charged n-gram scores, from most moderate/right at the top, to most liberal/left at the bottom

Candidate	Group
gottheimer_josh	Incumbent_Moderate
brindisi_anthony	Incumbent_Moderate
sherrill_mikie	Incumbent_Moderate
mcadams_ben	Incumbent_Moderate
lamb_conor	Incumbent_Moderate
murphy_stephanie	Incumbent_Moderate
lipinski_dan	Incumbent_Moderate
golden_jared	Incumbent_Moderate
spanberger_abigail	Incumbent_Moderate
luria_elaine	Incumbent_Moderate
fudge_marcia	Incumbent_Liberal
watsoncoleman_bonnie	Incumbent_Liberal
bass_karen	Incumbent_Liberal
gomez_jimmy	Incumbent_Liberal
grijalva_raul	Incumbent_Liberal
pressley_ayanna	Incumbent_Liberal
khanna_ro	Incumbent_Liberal
jayapal_pramila	Incumbent_Liberal
ocasiocortez_alexandria	Incumbent_Liberal
omar_ilhan	Incumbent_Liberal

To be sure, there's some circular logic with this set of members. The n-grams were initially chosen based on these members using or not using them, and then they were further weighted by how many members from each group either used them or didn't use them. However, this circular logic should only apply to the 20 benchmark members; the politically-charged n-gram score should be a perfectly reasonable measure for all other individuals in the pool.

Indeed, it does appear to "fix" some of the anomalies presented by the cosine similarity analysis. To stick with our earlier example, Marie Newman is now scored as the 13th most liberal individual from the entire pool using the politically-charged n-gram score. This means that she is now far to the "left" of Dan Lipinski, who is scored as the 11th most moderate individual using the politically-charged n-gram score. These scores line up much more closely with what you'd expect based on the conventional wisdom about these two individuals.

Conclusion

As the project is still a work in progress, there are a series of next steps that we plan to continue working on. First, we plan to increase the number of candidates and members by scraping additional campaign websites. Ideally, we would like to include the entire Democratic caucus in the House of Representatives. Second, we plan to explore the term frequency/inverse-document-frequency (tf-idf) method as a means of applying weighting to the cosine similarity analysis.[17] Third, we will continue to explore additional words and phrases that can contribute to the politically-charged n-grams analysis. Finally, as the ultimate goal is to have one single score that captures the ideology of a candidate/member, we will attempt to determine the optimal specification and weighting between the different measures in calculating a final ideology score.

While a significant body of work still remains to be done, the preliminary results of the analysis are promising. Further, while the initial goal of the project was to assess the ideology of the candidates in this particular election cycle, the methodology used in this analysis has the potential to be useful in assessing political ideology in a broader context as well. If the methodology can be further refined and validated, we hope that this type of analysis could serve as a supplemental measure to existing tools for measuring political ideology going forward.

---

[17] Bigi, Brigitte. "Using Kullback-Leibler Distance for Text Categorization." *Lecture Notes in Computer Science Advances in Information Retrieval*, 2003, pp. 305–319., doi:10.1007/3-540-36618-0_22.