1. Background

1.1. Introduction

Researchers Ludovic Rheault and Christopher Cochrane believed that by utilizing word embeddings augmented by political metadata, they could create a model that is better suited for political text than the typical ideological scaling methods. They argued that ideological scaling methods that measured word frequencies were not able to capture semantic properties of certain words to the same extent as word embeddings, which instead can use word vectors and neural networks to understand language in its full context. Furthermore, the addition of control variables, specifically party affiliation and date, accounts for government-opposition dynamics that previous researchers had not yet accounted for. Lastly, this method allows the researchers to situate political actors and parties based on their proximity to political concepts. They tested these claims on three large corpuses of political text from Britain, Canada, and the United States- ultimately finding that their "party embeddings" approach consistently aligns with human-annotated indicators such as voting history and expert surveys.

In our replication, we will be repeating the same preprocessing and model implementation steps to understand if their results are consistent across these three corpuses. Pulling the data from the three sources laid out in the research paper, we will first preprocess the speech transcripts, build a new model using the researcher's template, and validate using the golden standard documents distributed by the researchers.

1.2. Corpus details

1. **US House and Senate Corpus**: to analyze ideological placement within the U.S. Congress, the authors utilized parsed Congressional Records data as provided by Gentzkow, Shapiro and Taddy (2018). The corpus contains speech and speaker metadata from the 43rd to the 114th Congress, spanning from years 1873 to 2016. The speech data are taken from both the Congressional Records' bound and daily versions[1]; data for the 43rd and the 111th Congress are sourced from the Records' bound format, while the data for the 112th and 114th Congress are from the daily versions. Further, the metadata contains information on the speaker's Congress session, their first and last names, chamber, party affiliations, and their states.

2. **British Hansard Corpus**: The original corpus utilized by the authors was unavailable due to the deactivation of the original corpus' source project and data sharing restrictions. The alternative corpus utilized is published by Betto van Waarden and Mathias Johansson (2022) and mirrors the Rheault version(with permission) with additional metadata. Encompassing debates in both the House of Lords and House of Commons between 1935 and 2014, it comprises 3.4 million unique documents, with speech text and speaker metadata including the speaker's party affiliation, legislative session number and year of the legislative session.

3. **Canadian Hansard Corpus:** The entire corpus utilized by the researchers was made available through the Linked Parliamentary Data Project (LiPaD) website. This corpus, which was

[1] The bound versions are compilations of all daily versions, and are made up of one volume per Congress sessions. These differ from the daily versions mainly in regards to: (1) their continuous pagination features; (2) their somewhat revised, rearranged, and edited texts; (3) their removal of the prefixes H, S, and E before page numbers.

compiled by a multidisciplinary team of political scientists, computer scientists, and historians at the University of Toronto, contains a vast repository of files documenting every speech made in Canadian parliament from 1901-2019. Each file contains useful metadata on the date, speech topic, speaker party affiliation, and speaker name, as well as the transcript of the speech itself. The data can be downloaded in multiple formats, including a complete PostgreSQL dump or XML document, however for our purposes, we downloaded the full Canadian dataset in a series of daily CSV files.

1.3. Methods

All three corpora underwent identical pre-processing methods, as specified by the author in their published code specification. This involved removing: digits, words containing fewer than two letters, common English stop words, and unique sets of procedural words specific to each corpus. Additionally, the vocabulary was limited to words occurring more than 50 times in the corpus, and common phrases (bigrams and trigrams) were merged together, using the phraser function available through the Gensim Library - for eg, "New York" was transformed into "New_York." Further, two document level tags were created for each document. The first tag indicated the party of the speaker and legislative session number dyad. The second tag indicated the legislative session number.

The Doc2Vec models trained for all three corpora utilized identical hyperparameters and training specifications, as detailed by the author. These specifications included a learning rate of 0.025, a training duration of 5 epochs, a model architecture with 200 hidden layers, and a context window size of 20. Once trained, the models yielded embeddings for the document level tags described above indicating unique semantic and thematic information about each party during a legislative session. We next performed Principal Components Analysis on the document embeddings, interpreting the 1st and 2nd PCs as the ideological placement of parties and the most important source of semantic variation across parties in a legislature, respectively.

Finally, we validate the results of our embeddings by adhering to the author's methodology of using externally valid, gold-standard measures of party ideology to assess the correlation and pairwise accuracy of our first principal component against the gold-standard measures. These gold standard measures include Voteview scores, expert survey assessments, as well as three measures that are derived from the Comparative Manifestos Project (CMP), namely Rile, Vanilla, and Legacy[2]. Additionally, a WordFish model is also trained on the US House corpus. The authors then compare the correlation and pairwise accuracy of the Voteview gold-standard measure with both the ideological placements obtained from the WordFish model and those from the Doc2Vec embeddings for the US House corpus.

Lastly, the authors also demonstrate the usefulness of their approach for measuring legislators' ideological placements. In this exercise, they include speaker metadata as part of the document tags in the Doc2Vec training process, with the aim of producing speaker-level embedding which can then be interpreted as their ideological placements. They show the validity of their results by comparing their measures with known gold standard measures for identifying legislator-level political leanings: DW-Nominate, Nokken-Poole, ACU Life, ACU 2015-2016, and GovTrack scores.

[2] Detailed explanations on each of these measures can be found in pages 124-125 of the original paper.

2. Replication Results

We then conduct the corpus pre-processing and model training with guidance from the scripts which the authors have provided in their GitHub repository. After the models on each corpus are trained, we proceed to fit a two-component PCA model on the resulting party-level embeddings to identify the ideological placement of the US, British and Canadian parties.

Figure 1 shows the scatterplot of the PCA results from our trained Doc2Vec model and the authors' original results for the US House corpus. The figure suggests that our replication results closely resemble the pattern observed in the original results, with minor differences in the location of the dot points attributable to differences in random initialization of weights during the Doc2Vec training process. Further, we are also able to closely replicate the authors' ordering of the sessions. For example, consistent with the original results, we also locate Dem-2011 in the top left quadrant and Rep 2011-2015 in the bottom right quadrant.

To interpret these two principal components, the authors look at the words that are closest to the edges of these PCs' spaces. The results from our replication of this exercise are shown by Table 1. Although there are some differences in words that appear in the edges of each PC axes, we find that the words in each edge still form relatively similar themes as compared to those in the original paper. For the right edges of the first PC, we still identify words such as "bureaucrat", "bureaucratic", "communist"–words that characterize issues that are relevant to the right-leaning Republican party. On the other hand, words on the left edge of the first PC include "Congressional Black Caucus", "racism", "African American", "civil rights", all of which are issues that are relevant to the Democratic party. Further, as with the original paper, the edges of the second PC include words that differentiate between the North-South axis.

With the above interpretation of PCs as a backdrop, the authors then plot time series charts for each party to identify party-level trends for the two PCA components. Figure 2 and Figure 3 also show that our replication results closely follow the results that the authors show in their paper. In Figure 2, both results virtually show a widening gap in the first principal component value, which indicates growing divergence in ideological placement that is reflected by word use patterns that have become increasingly divergent over time. We are also able to reproduce the reversal trend of the second PC that is observed in Figure 3 panel (b). Such a high degree of similarity between the results from our replication and the authors' original ones is likely attributable to the fact that we are using the same version of the corpus that the authors were using. Further, the availability of the pre-processing scripts that are specifically tailored to the US corpus also helps identify the corpus-specific pre-processing decisions and greatly aids the replicability.

In replicating the ideological placement for the British Corpus, similar to the analysis of the US corpus, we plot the party embeddings gained from our British Doc2vec model (see Figure 4). Here, we find that party placements broadly align with the ideological positions identified by the original authors. Specifically, the Labour Party is positioned on the left, indicating liberal leanings, the Liberal Democrats are in the center, and the Conservative Party is on the right, reflecting conservative leanings. Our embeddings capture major trends, such as positioning members of the 1987 Conservative Party as the most conservative, which aligns with the widely recognized highly conservative stance of Margaret Thatcher's government (also aligning with the author's findings). However at a finer level, our results show higher variance in party placements, especially for the Conservative Party, which, in several years, appears closer to the Labour Party and generally spans across the ideological scale, contrary to what the author's original findings suggest, where party

positions are distinctly separated by their ideology. Additionally, our replication shows both the Labour and Conservative Parties for certain years gravitating more towards the ideological center.

This discrepancy likely stems from differences in the corpus used and the pre-processing methodologies between our study and the original. The authors removed speeches from the chamber's Speakers which are primarily neutral and bipartisan due to the Speaker's role. Our dataset lacked indicators to identify such speeches, preventing us from taking explicit steps to ensure exclusion. Thus there is ambiguity on whether the British Corpus used in this replication has, at its source, speaker speech removed or not. If our embeddings are capturing the chamber speaker's speech as well, which is generally bipartisan and neutral, along with the MPs' it can cause the document embeddings for different party-years to appear more similar, as we see in our replication results. Secondly, while the authors detailed their pre-processing steps for the US corpus, specifics for the British Corpus were absent. We thus adapted and emulated the US corpus's pre-processing methodology for the British corpus. In doing so we may have missed specific pre-processing steps the author's took, unique to the British Corpus, to de-noise the documents effectively. For e.g. in the study they note that they manually correct the party affiliations for certain members who flipped between parties over the years, but do not provide any information regarding replication of this step. Corpus specific pre-processing may thus have yielded the more separable and less noisy embeddings in their results. Without replicating the exact pre-processing methods undertaken for this corpus, our model thus may be capturing a lot of noise as well as we see in our results.

[Grace to include discussion (and differences) on Canada]

We then proceed to replicate the results of the validation exercise that compares the party-embedding measures with the selected gold standard metrics. These are shown by Table 2. In general, compared to the original results, our trained models have similar performance, or even better in several metrics, when trained on the US corpus. For the US House corpus, for example, our trained model has 0.93 correlation with the Voteview measure, compared to 0.918 from the original paper. It also has somewhat higher accuracy; our model is able to achieve 100% pairwise accuracy with the expert surveys, compared to the original paper's 86.67% on the same metric. However, for the UK and the Canadian corpus, our models have weaker correlations and lower pairwise accuracies with the gold standard measures as compared to the authors' models. For the UK corpus, the correlation between our embedding measure and the expert surveys' measure is substantially lower (0.717) compared to the authors' results (0.910). It also has lower pairwise accuracy (77.78%, compared to the authors' 83.33%). Similarly, for the Canadian corpus, we are only able to achieve 0.188 correlation with the Legacy gold standard measure, as compared to the authors' correlation of 0.855. The lower accuracy and weaker correlation on the UK and Canadian corpus likely stem from the issues described above.

[Ayush to include discussion on the wordfish validations]

[Grace to include discussion on legislator-embeddings]

3. Autopsy

Our group ran into several issues during this replication process due to a lack of visibility around each of the preprocessing steps. While the researchers did provide ample descriptions for the United States corpus, we were tasked with adapting the preprocessing and word embedding code

for our Canadian/British implementation, despite notable discrepancies in the datasets. For example, in the Canadian corpus, there is no variable describing the session number associated with each speech (which is later used during model implementation), meaning we needed to manually map the year from the data column to a dictionary containing corresponding year ranges. Additionally, there was a lack of standardization around the party names, meaning that while the researchers may say in the final paper that they only included text from the 'conservative' or 'NDP' party, it's unclear to us whether that includes 'conservative (1867-1942)' or 'Independent Conservative', 'National Democratic Party', etc.

Furthermore, there were some issues with backward compatibility, seeing as

4. Extension

There are several possible extensions to the paper's analysis. First, in their paper, the authors are exploring the *overall* ideological placement along party lines. It would be interesting to further this analysis by exploring variations in party-level ideological placement across different issue topics, such as immigration, abortion, economic redistribution, among others. Second, we think it is also important to not only present descriptive trends in ideological placement divergences, but also to incorporate uncertainty and sampling variations in the analytical framework by measuring the statistical significance of such placement divergences (Rodriguez, Spirling and Stewart, 2023). Third, another potentially useful approach for exploring party-level ideological placement and polarization is by applying Kozlowski, Taddy and Evans (2019)'s geometry of embedding approach to situate political parties and legislators on a predefined political issue spectrum. It can also serve as a validation check to the authors' interpretation of the two principal components axes. Further, with this approach, one would also be able to identify legislators' positions along certain political issues. Lastly, it would also be fruitful to explore whether similar ideological placement accuracy can be achieved using other sources of political texts, such as Senators' tweets, news outlets, among others.

5. References

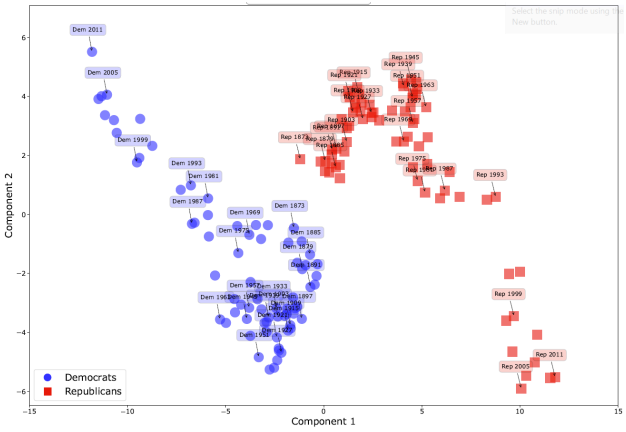
Gentzkow, M., Shapiro, J. M., & Taddy, M. (2018, January). Congressional record for the 43rd-114th congresses: Parsed speeches and phrase counts. In *URL: https://data. stanford. edu/congress text*.

Kozlowski, A. C., Taddy, M., & Evans, J. A. (2019). The geometry of culture: Analyzing the meanings of class through word embeddings. *American Sociological Review*, *84*(5), 905-949.

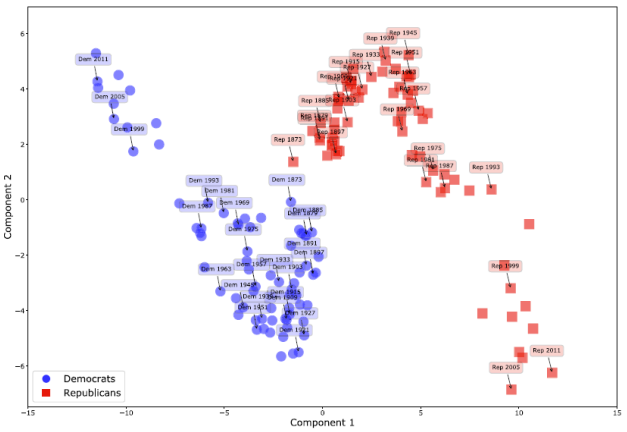
Rheault, L., & Cochrane, C. (2020). Word embeddings for the analysis of ideological placement in parliamentary corpora. *Political Analysis*, *28*(1), 112-133.

Rodriguez, P. L., Spirling, A., & Stewart, B. M. (2023). Embedding regression: Models for context-specific description and inference. *American Political Science Review*, *117*(4), 1255-1274.

Figures and Tables

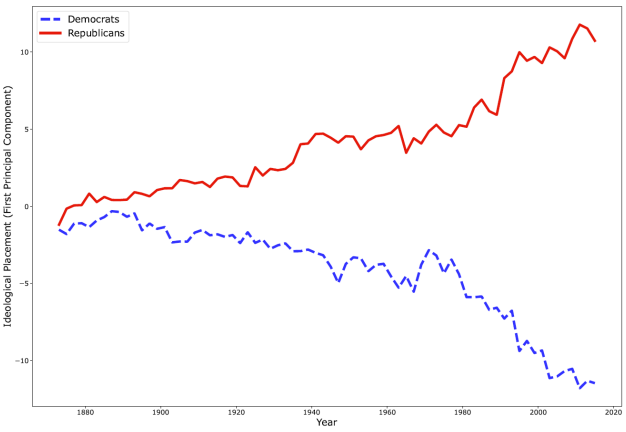


(a) Replication

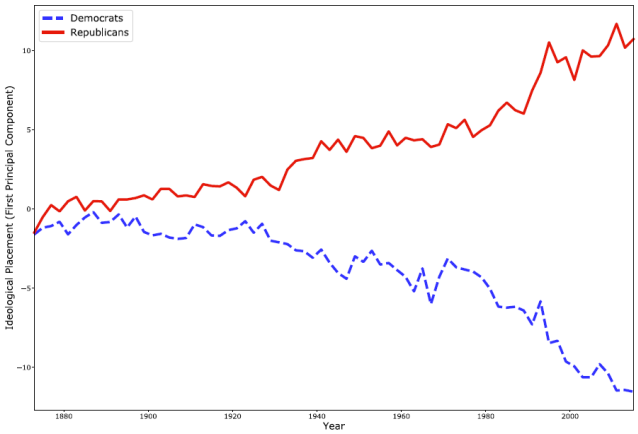


(b) Original

Figure 1. Party Placements in the U.S. House

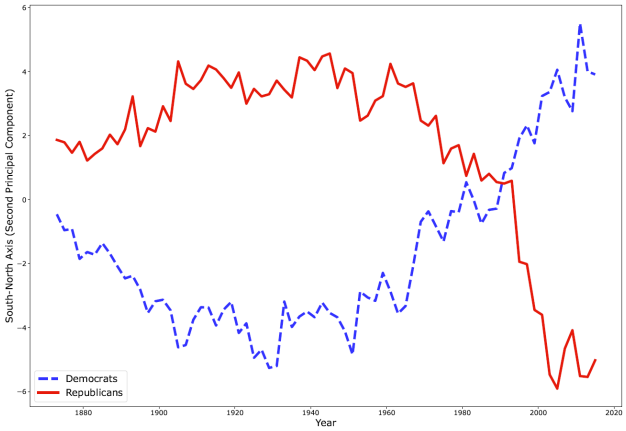


(a) Replication

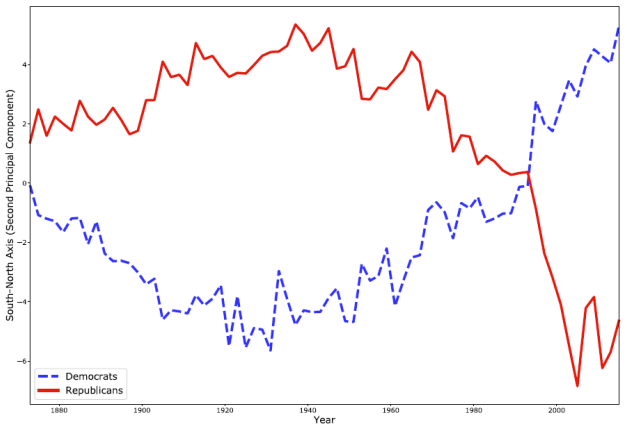


(b) Original

Figure 2. First PCA Component across the years



(a) Replication



(b) Original

Figure 3. Second PCA Component across the years

Table 1. Interpretation of the PCA Components

Component	Orientation	Words/phrases closest to the edges of the PCA space
First	Positive (right)	Nebraska, Obamacare, savings account, socialized medicine, bureaucracies, bureaucracy, Missouri river, bureaucratic, wheat, feed grains, South Dakota, feed grain, socialism, forest reserves, free enterprise, IRS, redtape, Yugoslavia, savings accounts, hogs
	Negative (left)	wealthiest, South African, Congressional Black Caucus, racism, decent housing, civil rights, voting rights, civil rights movement, South Africa, poor elderly, message announced, infant mortality, millionaires, African Americans, Blacks, impoverished, CBC, rich poor, segregated, Chile
Second	Positive (North)	City Detroit, Mich, Toledo, Seattle, plant closing, balance, Hartford, Rochester, Akron, Chrysler, Northern Ireland, free trade, West Germany, TAA, Western New York, layoff, retraining, Vermont, Duluth, Niagara Falls
	Negative (South)	Oklahoma, Everglades, Georgians, Fort Benning, textile imports, parish, Georgian, civilized tribes, court house, Gainesville, Fort Smith, long staple cotton, judge, tyrant, Savannah river, Shreveport, Southeast, flue cured tobacco, happiness, Baptist church

Table 2. Accuracy of Party Placement as Validated against Gold Standards

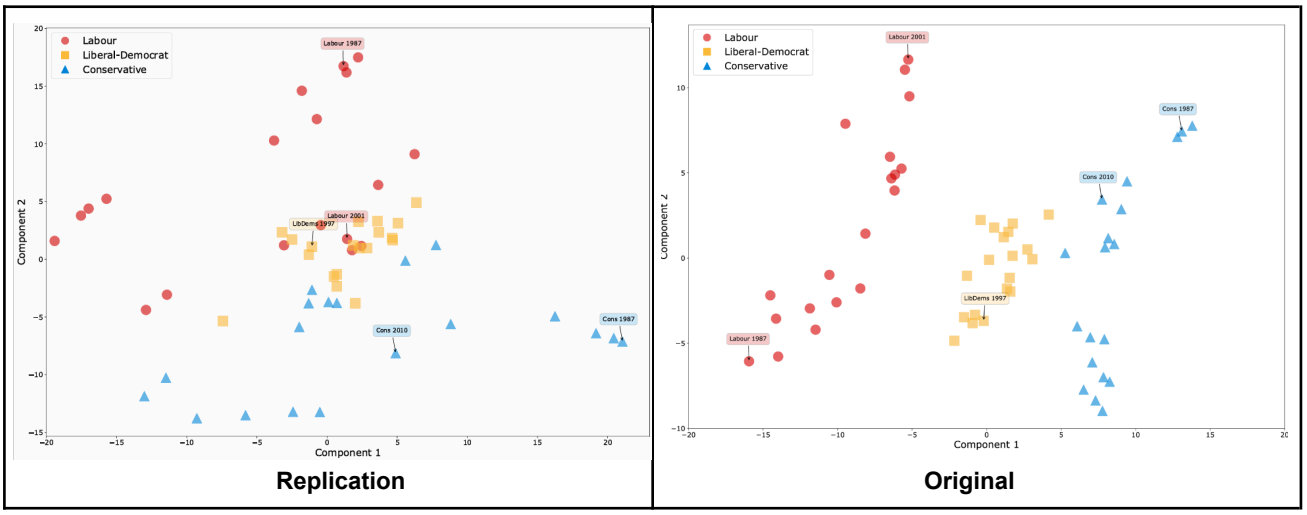
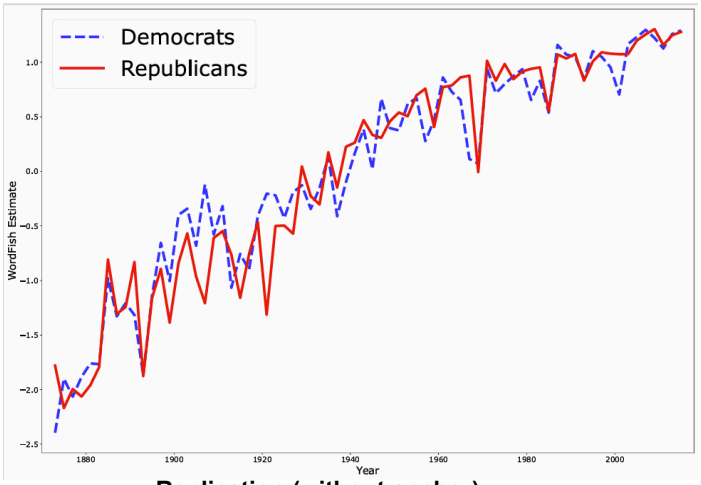


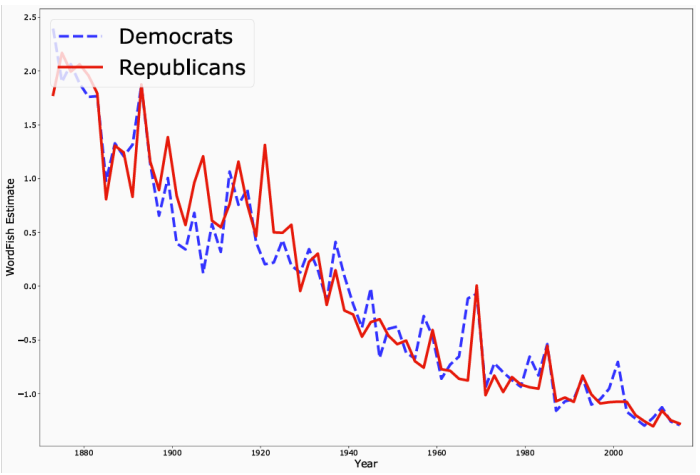
Figure 4. Party Placement in the British Parliament

Table 2. Accuracy of Party Placements Validated against Gold Standards

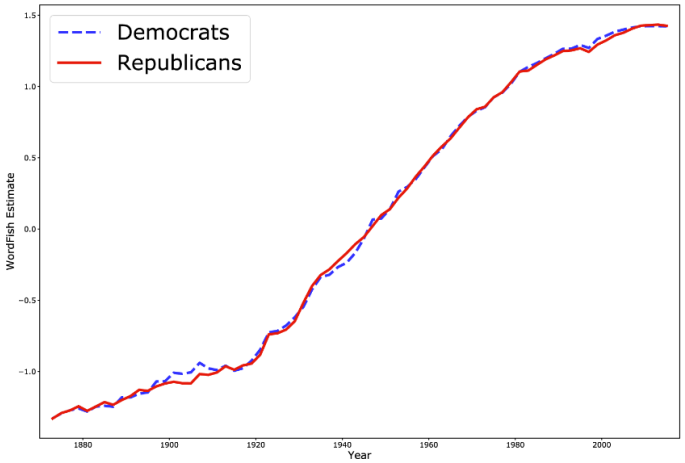
Gold standard	Metric	US House	US Senate	Canada	Britain
Voteview	Correlation	0.930	0.890		
(1921-2016)	Pairwise accuracy	85.83%	83.51%		
Expert surveys	Correlation	0.988	0.986	0.029	0.717
(1984-2002)	Pairwise accuracy	100%	100%	52.27%	77.78%
Rile	Correlation	0.631	0.628	0.178	0.503
(1945-2015)	Pairwise accuracy	72.96%	72.07%	53.44%	61.43%
Vanilla	Correlation	0.742	0.722	0.021	0.493
(1945-2015)	Pairwise accuracy	76.15%	74.73%	50.67%	58.42%
Legacy	Correlation	0.899	0.890	0.188	0.360
(1945-2015)	Pairwise accuracy	85.55%	85.02%	51.05%	58.70%



Replication (without anchor)

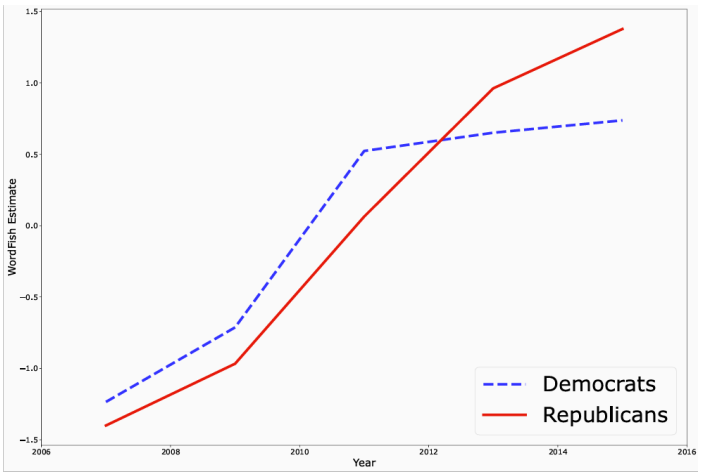


Replication (with anchor)

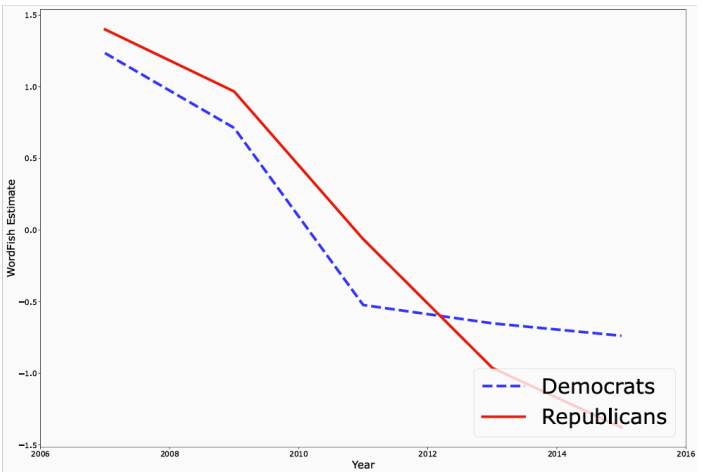


Original

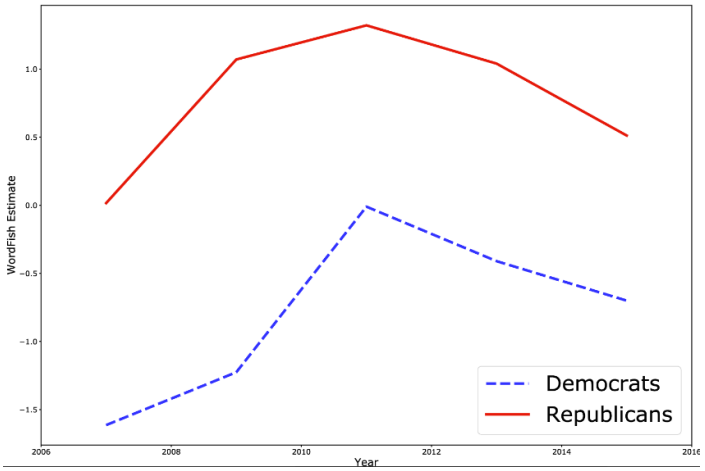
Figure x. Wordfish Estimates: US House (1873-2016)



Replication (without anchor)



Replication (with anchor)

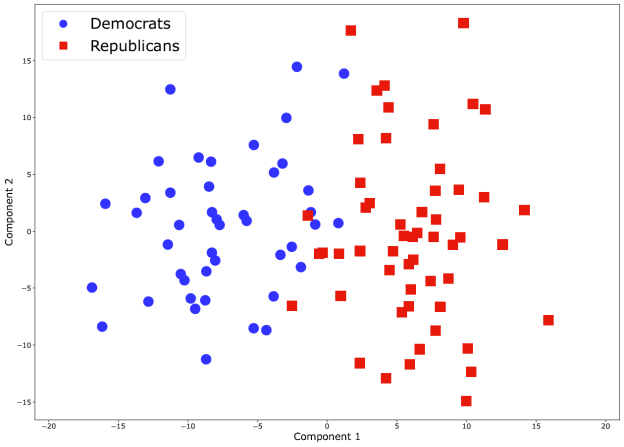


Original

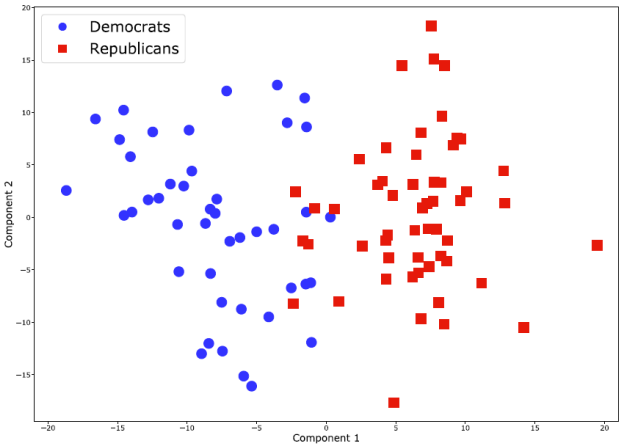
Figure x. Wordfish Estimates: US House (2007-2016)

Metric	1921–2016		2007–2016	
	Wordfish	Embeddings	Wordfish	Embeddings
Correlation	-0.032	0.854	-0.023	0.874
	(-0.027)	(0.918)	(0.829)	(0.999)
Pairwise Accuracy	52.3%	61.21%	44.44%	62.22%
	(44.36%)	(85.66%)	(75.56%)	(88.89%)

Table 3: Comparison of WordFish Estimates and US House Embeddings (Parenthesis indicates author's results)



(a) Replication



(b) Original