PPOL 6801

Word Embeddings for the Analysis of Ideological Placement in Parliamentary Corpora

GRACE JENSEN, AYUSH LAHIRI, SEAN HAMBALI

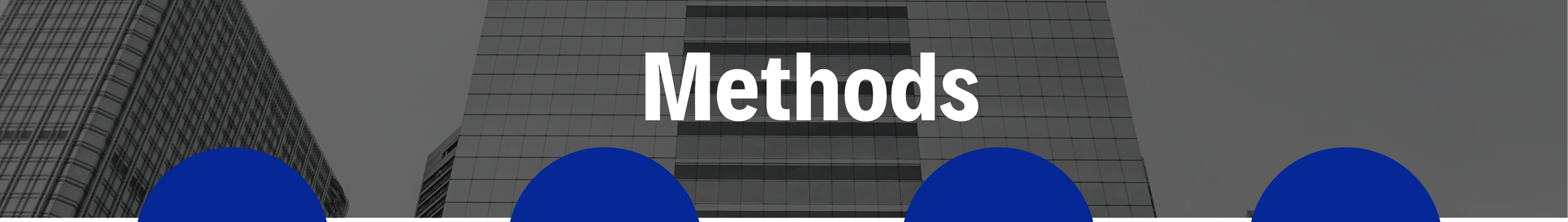
BACKGROUND

Ludovic Rheault and
Christopher Cochrane (2016)

‘Party Embeddings’

Wanted to create a word embedding model that would evaluate the language’s context, accomodate for different control variables, and situate actors based on their proximity to political concepts

Researchers enhanced word embedding models using metadata from political text- specifically party and date



Methods

Data

- Parliamentary debates from the United States, Britain, and Canada
- Focused on speeches from the major parties over the past century

Preprocessing

- Removed digits, words with <2 letters, English stop words, and unique sets of procedural words for each corpus
- Limited vocabulary with < 50 occurrences
- Merged common phrases together

(New York → New_York)

Fitting the Model

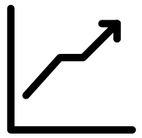
- Learning rate = 0.025
- Epochs = 5
- Hidden layers = 200
- Window size = 20
- Party-specific indicator variables
- Fit PCA from embedding to extract political position

Validation

- Compared predicted ideological placement against gold standards for each corpus
- Calculated correlation and pairwise accuracy

Corpus Details: U.S. Congress

Source: Gentzkow, Shapiro, Taddy (2018)'s parsed Congressional Record data



Covers 43-114th Congress (1873-2016)

Available in Bound and Daily versions

Includes speech and speaker metadata

Congress session	Speaker ID	Speaker Name	Chamber	Party	...	Original Text	Processed Text
43	43044451	Hannibal Hamlin	Senate	R	...	I move that nntil otherwise ordered the hour of the daily meeting ...	nntil ordered hour daily meeting ...

Corpus Details: British Hansard

Source: Betto van Waarden, Mathias Johansson (2022) corpus based on the Rheault version reproduced with permission along with additional metadata

Covers House of Lords and House of Commons (1935 - 2014, 3.4 million unique documents) with speech text and speaker metadata.

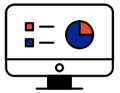
Differences: No identification of documents for speakers, no manual correction in party affiliation

Year	Parliament	Party	Original Text	Processed Text
1987	50	Lab	Will the Secretary of State issue an instruction to all regional health authorities not to close operating theatres over Christmas, as this would enable a further reduction to be made in the waiting lists?	state issue instruction regional health authorities close operating theatres christmas enable reduction waiting lists

Corpus Details: Canadian Hansard

Source: Lipad database via University of Toronto researchers

Series of daily CSV files that required additional preprocessing

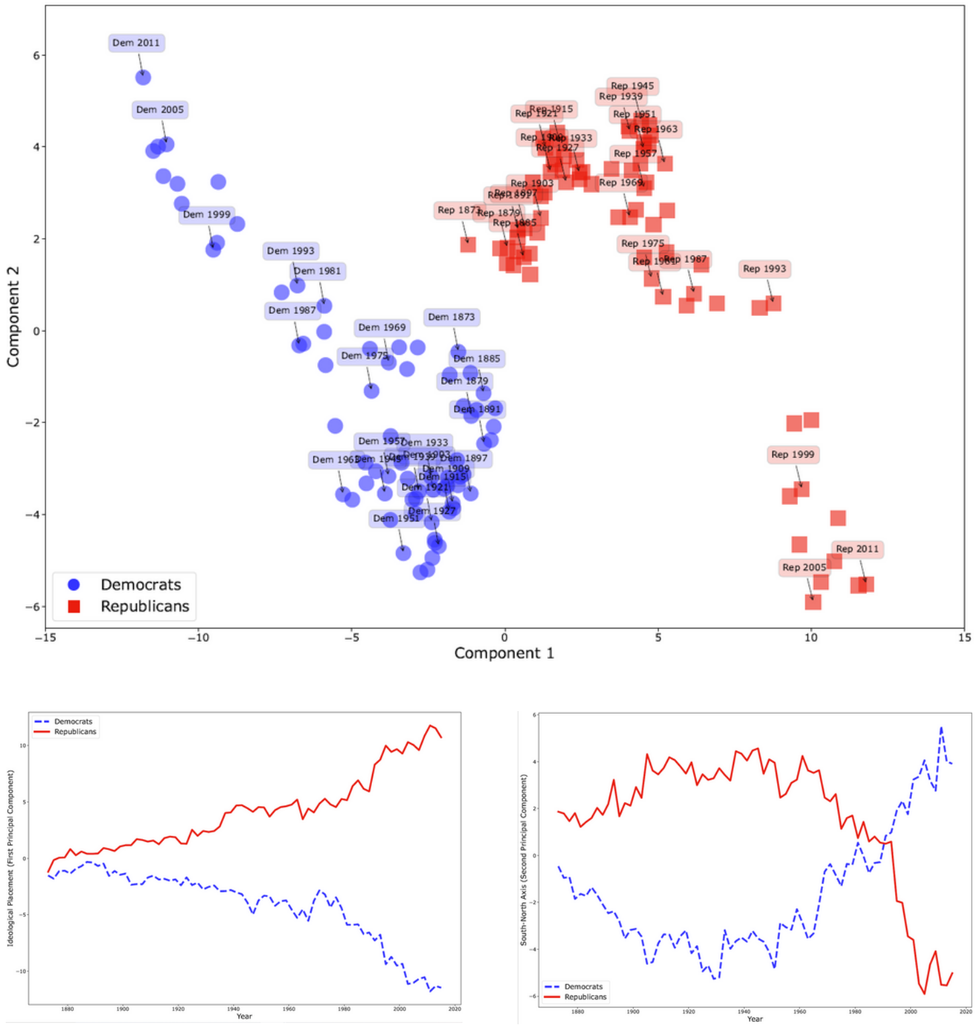


Covers Parliamentary debates from 1901-2019

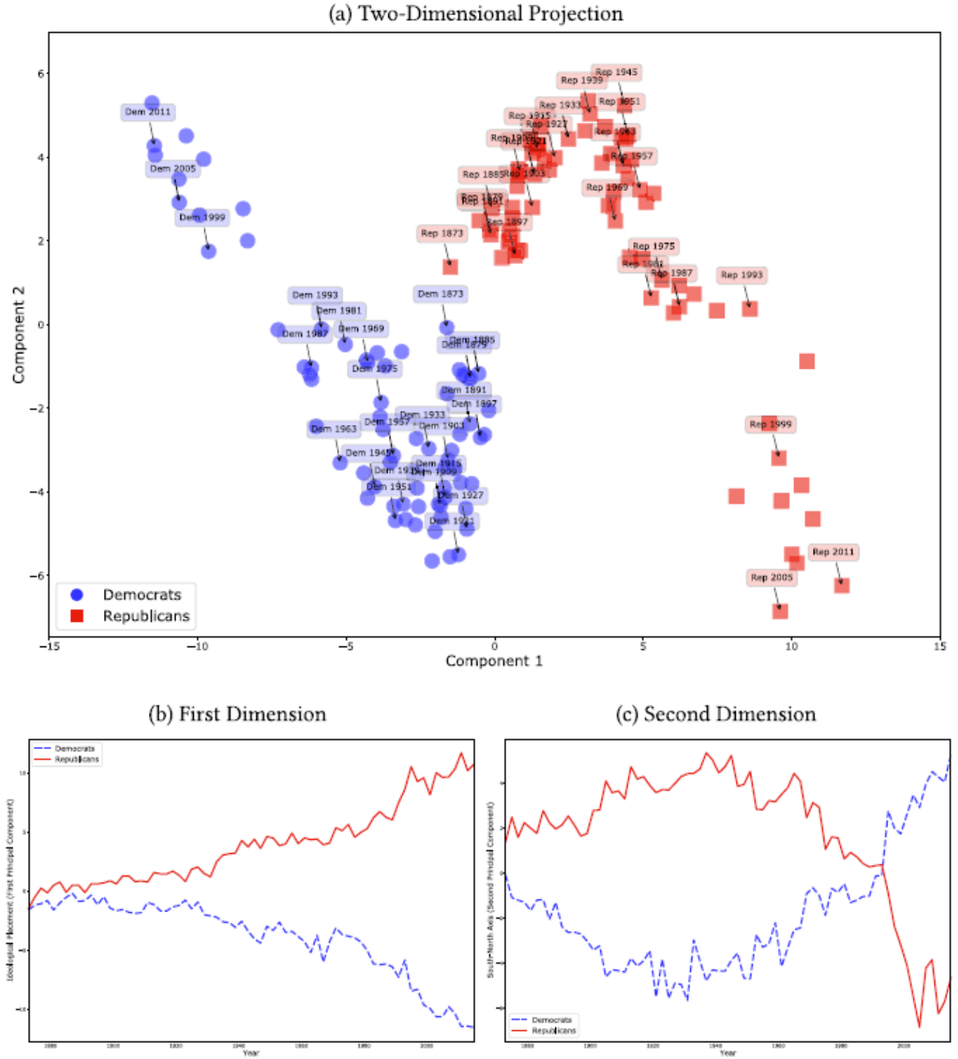
Differences: No session number, lack of standardization around party affiliations

speechdate	maintopic	...	speakerparty	speakername	speechtext	preprocessed text
2004-05-12	Statements by Members		LIberal	Dale Johnson	“Mr. Speaker, May 20 will mark the first anniversary of the discovery of BSE...”	“mark anniversary discovery bse”

Figure 2: Party Placements in the U.S House



Replication Results: House Corpus



Author's Results: House Corpus

Table 1: Interpreting the PCA Axes

Component	Orientation	Words and phrases
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, africanamericans, blacks, impoverished, cbc, rich poor, segregated, chile
Second	Positive (North)	city detroit, mich, toledo, seattle, plant closing, balance, hartford, rochester, akron, chrysler, northern ireland, freetrade, 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, courthouse, gainesville, fort smith, longstaple cotton, judge, tyrant, savannah river, shreveport, southeast, fluecured tobacco, happiness, baptist church

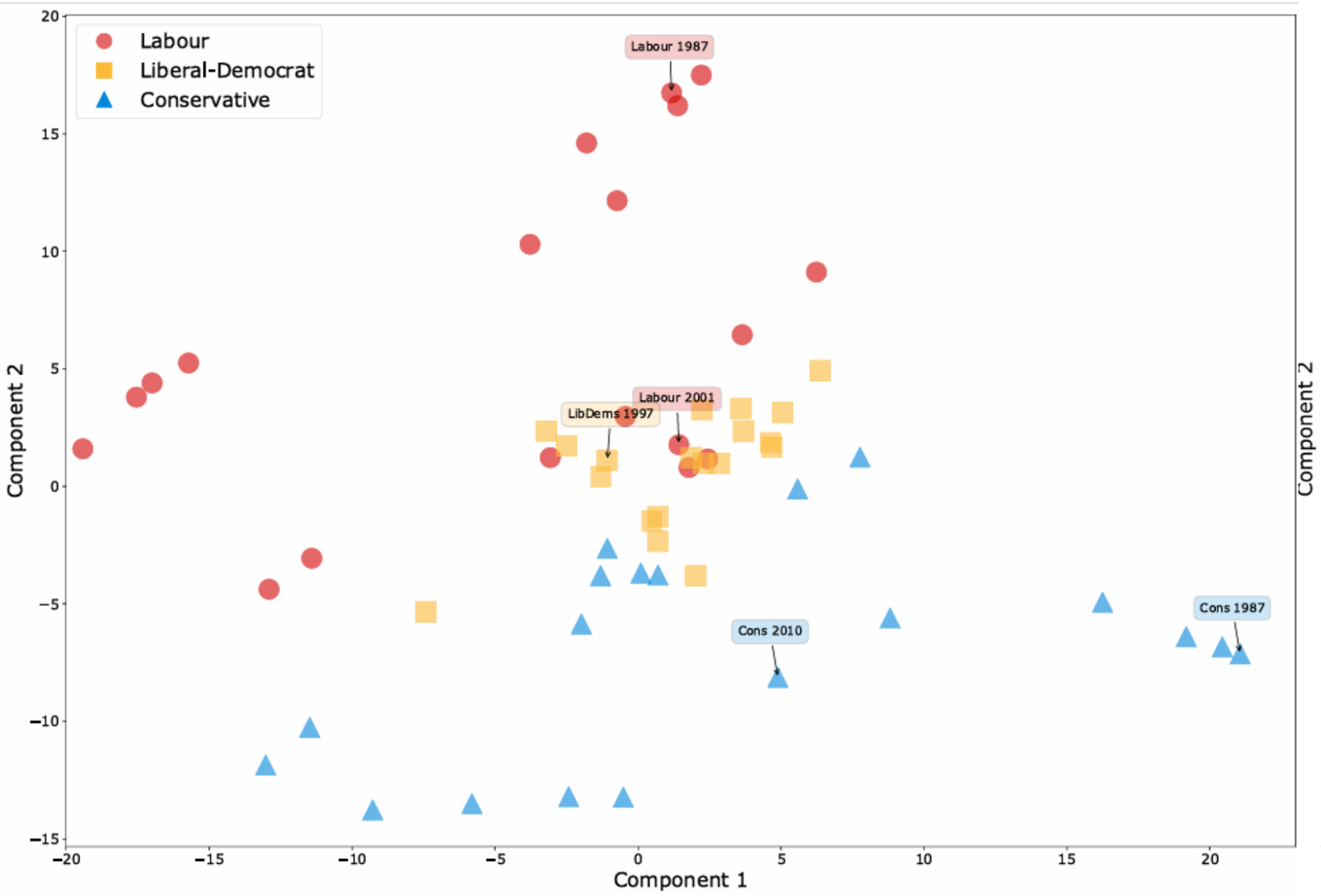
Replication Results: House Corpus

Component	Orientation	Words and phrases closest to edges of the PCA space
First	Positive (right)	Bureaucracy, bureaucracies, bureaucratic, Nebraska, regimentation, bureaucrats, Missouri River, Obamacare, centralized, redtape, Kansas, Hatch Act, charter schools, captive nations, free enterprise, Hoover Commission, Lenin, communist, feed grains, wheat
	Negative (left)	Congressional Black Caucus, wealthiest, decent housing, South African, slums, racism, poor elderly, African Americans, Latinos, African American, segregated, civil rights, gun violence, apartheid, African, poorest, joint resolution res, tax breaks, Brooklyn, richest
Second	Positive (North)	Detroit, Buffalo, Seattle, Minneapolis, Vermont, Duluth, Rochester, Lake Erie, debt gratitude, Lake Michigan, Erie, Cleveland, cleanup, retraining, Toledo, Maine, trade adjustment assistance, Oregon, Chicago, recycling
	Negative (South)	Brooks, usury, Sam Houston, parishes, Rome, cotton acreage, Poage, tobacco growers, Sam, burley tobacco, tyrannical, South Carolina, pink bollworm, southwest, Athens, fertilizer, military installations, North Carolina, Andrew Jackson, monopolies

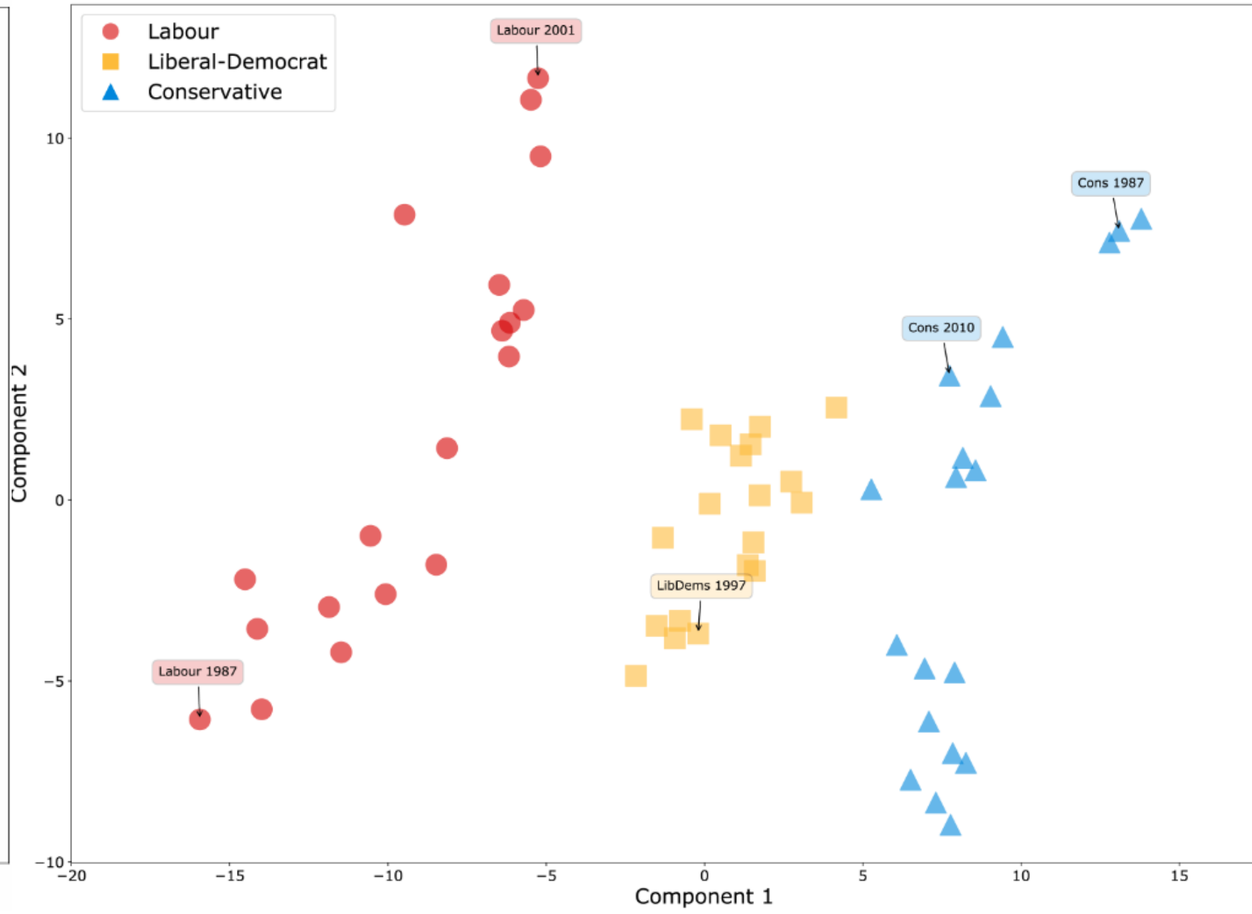
Author's Results: House Corpus

Figure 3A. Party Placement in the British Parliament

1st and 2nd Principal Components of Word Embeddings



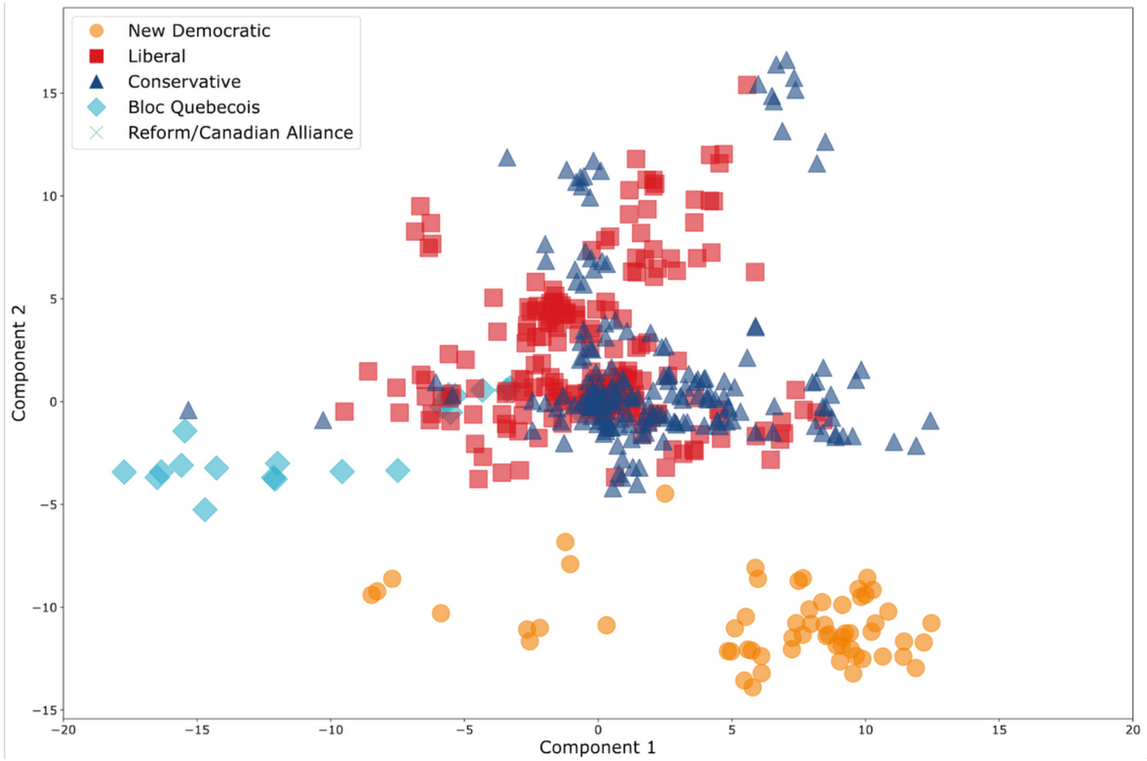
Replication Results: British Corpus

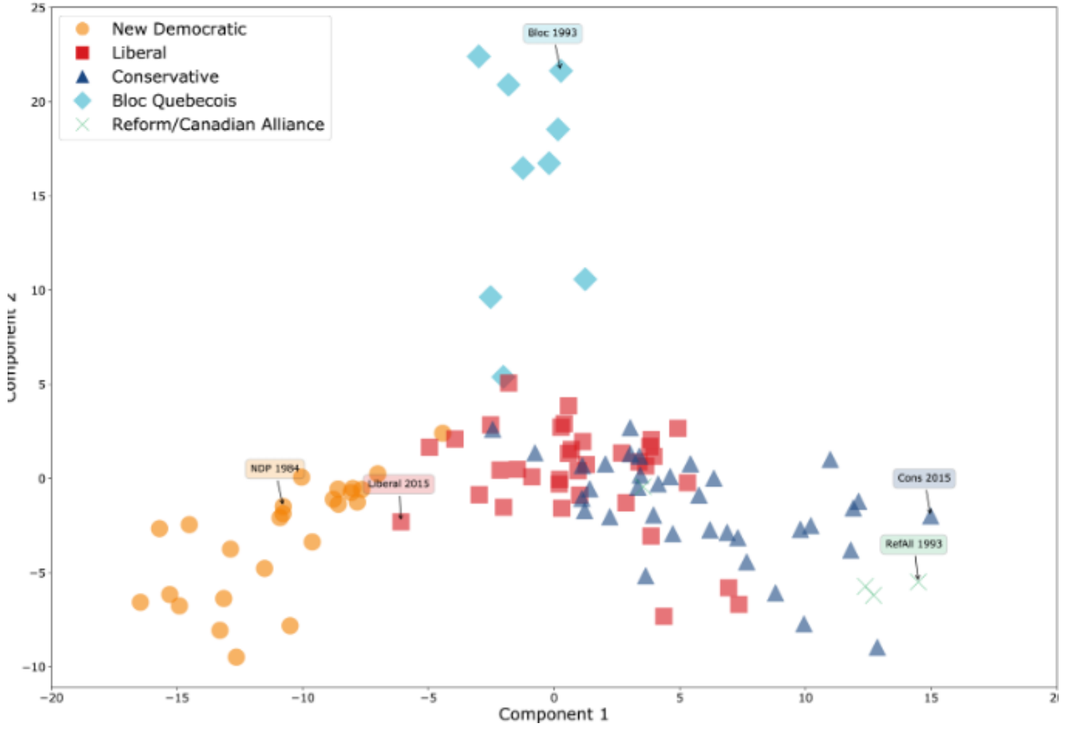


Author's Results: British Corpus

Figure 3B. Party Placement in the Canadian Parliament

1st and 2nd Principal Components of Word Embeddings





Replication Results: British Corpus

Author's Results: British Corpus

Table 2. Accuracy of Party Placement against Gold Standards

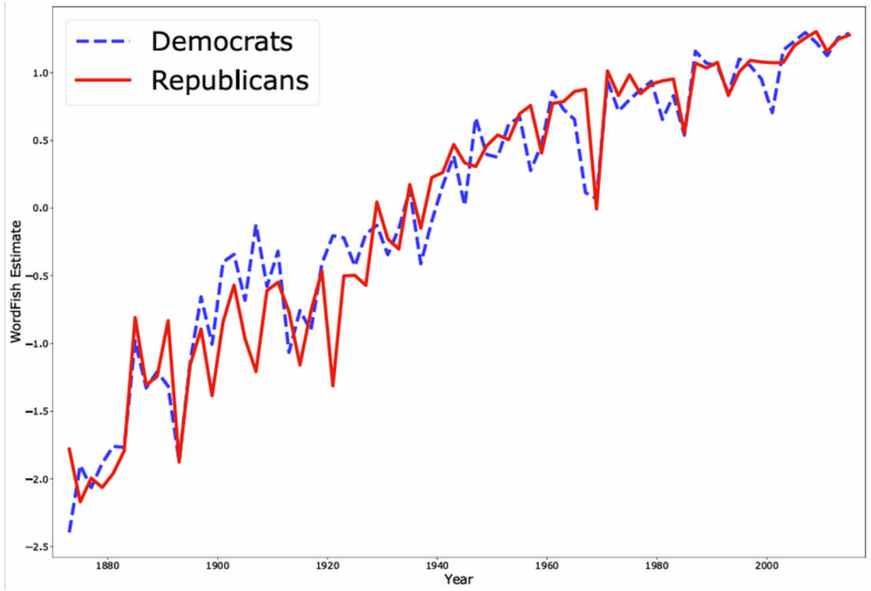
Gold standard	Metric	US House	US Senate	Canada	Britain
Voteview (1921–2016)	Correlation	0.918	0.919		
	Pairwise accuracy	85.66%	83.93%		
Expert surveys (1984–2002)	Correlation	0.982	0.981	0.869	0.910
	Pairwise accuracy	86.67%	100.00%	87.88%	83.33%
Rile (1945–2015)	Correlation	0.624	0.634	0.768	0.678
	Pairwise accuracy	72.61%	72.25%	77.03%	74.98%
Vanilla (1945–2015)	Correlation	0.736	0.736	0.731	0.755
	Pairwise accuracy	75.62%	75.09%	76.30%	78.06%
Legacy (1945–2015)	Correlation	0.898	0.907	0.855	0.876
	Pairwise accuracy	85.55%	85.37%	79.78%	82.67%

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

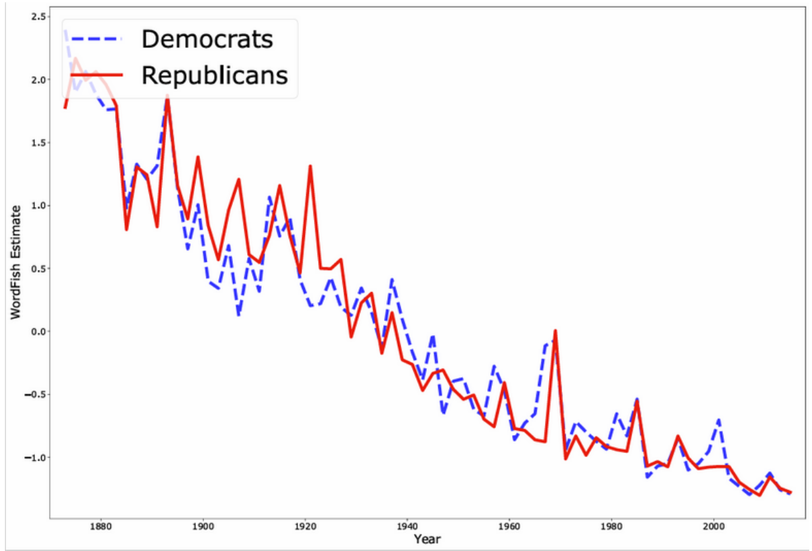
Figure 4. Wordfish Estimates: US House (1873–2016)

Replication Results: Full Corpus

Without Anchor



With Anchor



Author's Results: Full Corpus

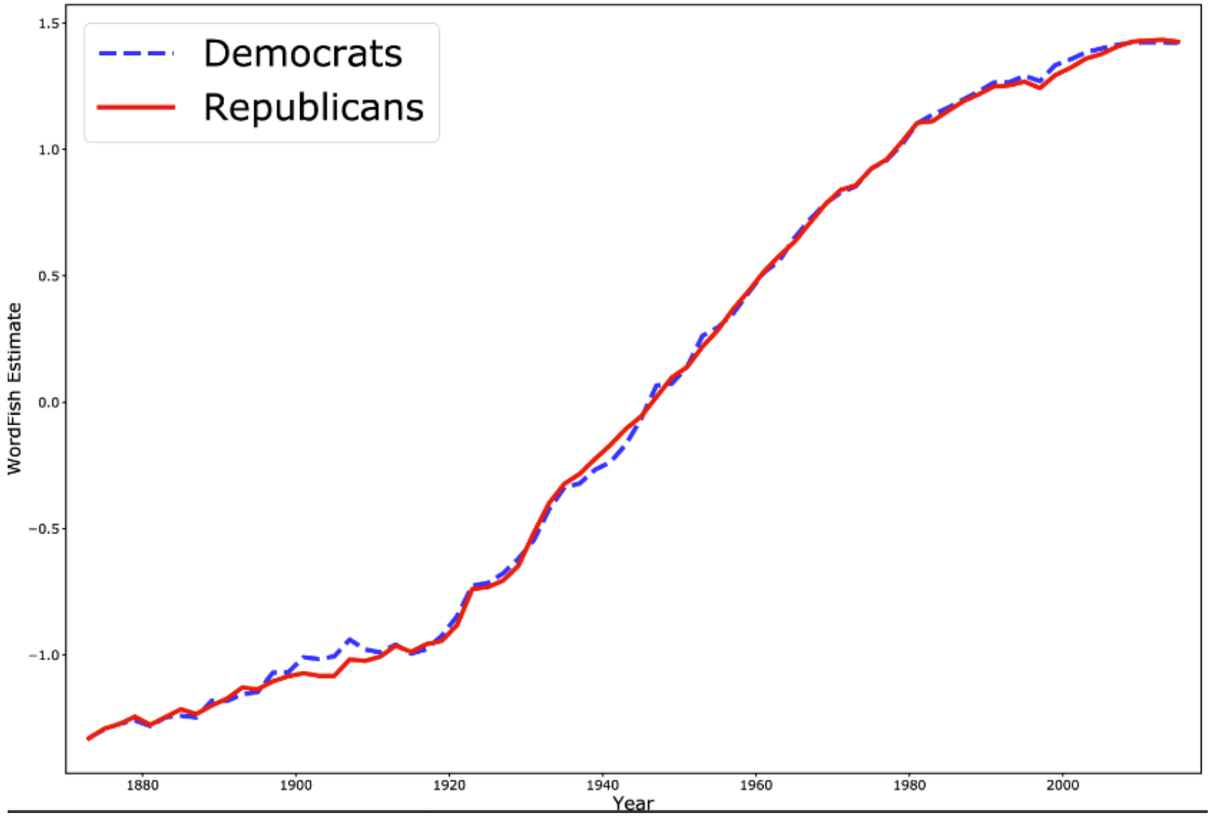
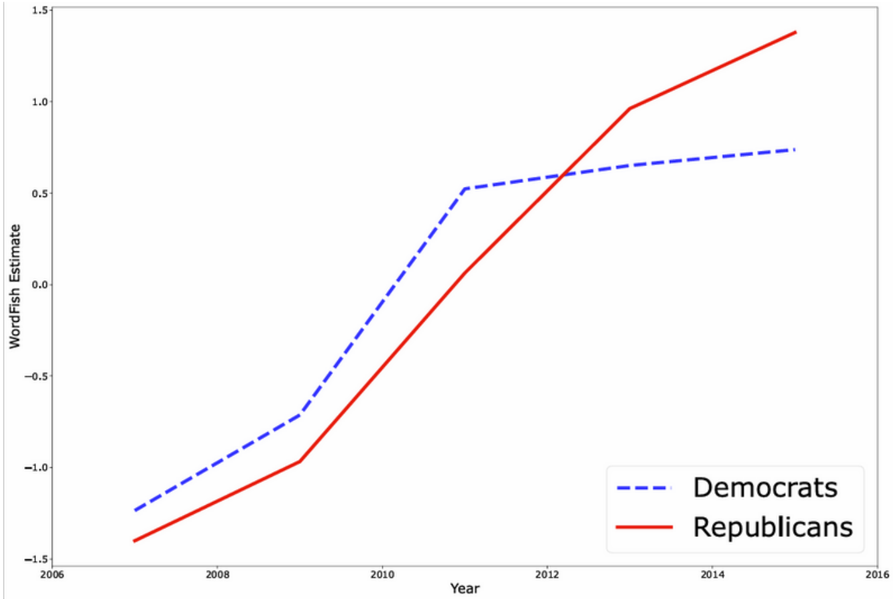


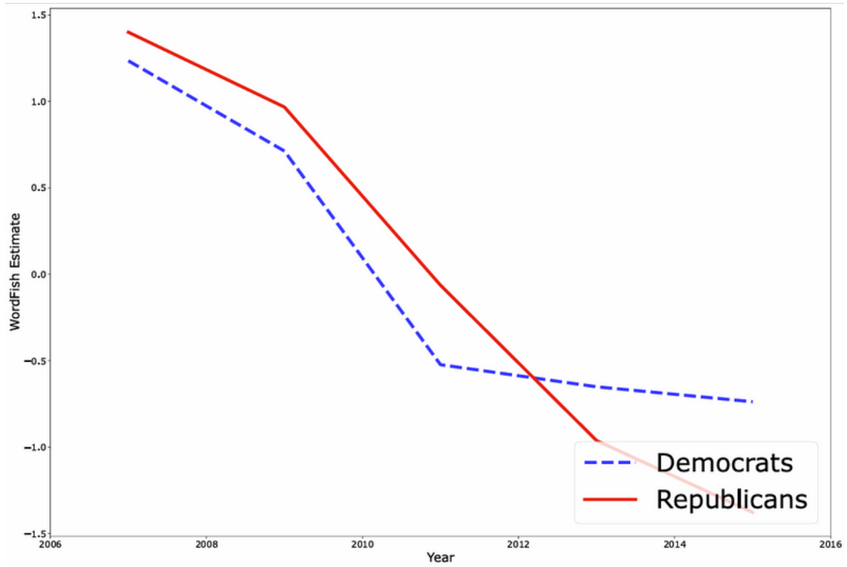
Figure 4. Wordfish Estimates: US House (2007–2016)

Replication Results: Most Recent Congresses

Without Anchor



With Anchor



Author's Results: Most Recent Congresses

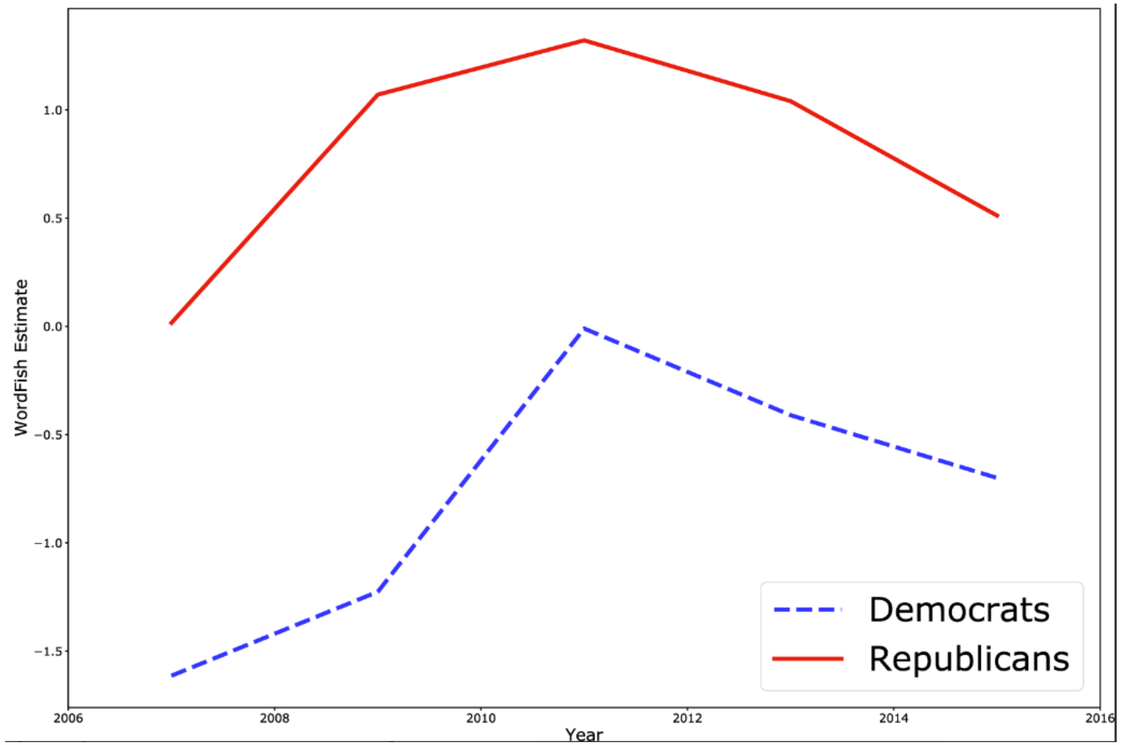
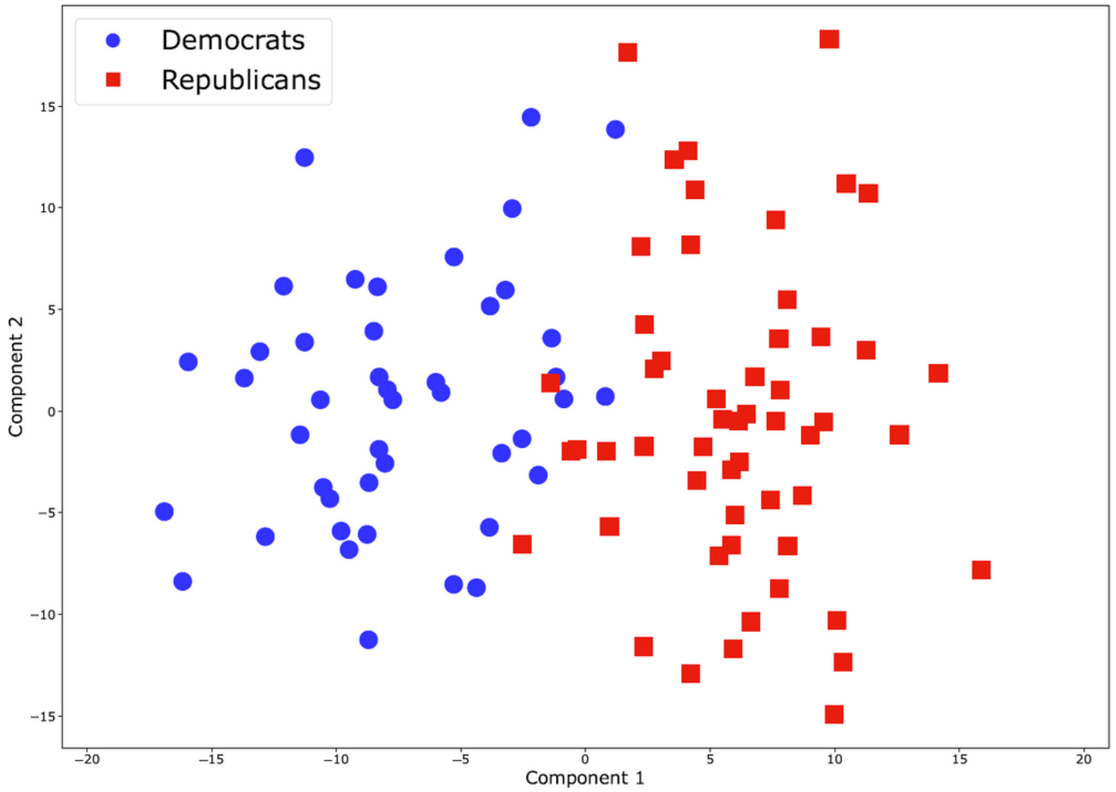


Table 3. Comparison with WordFish Estimates

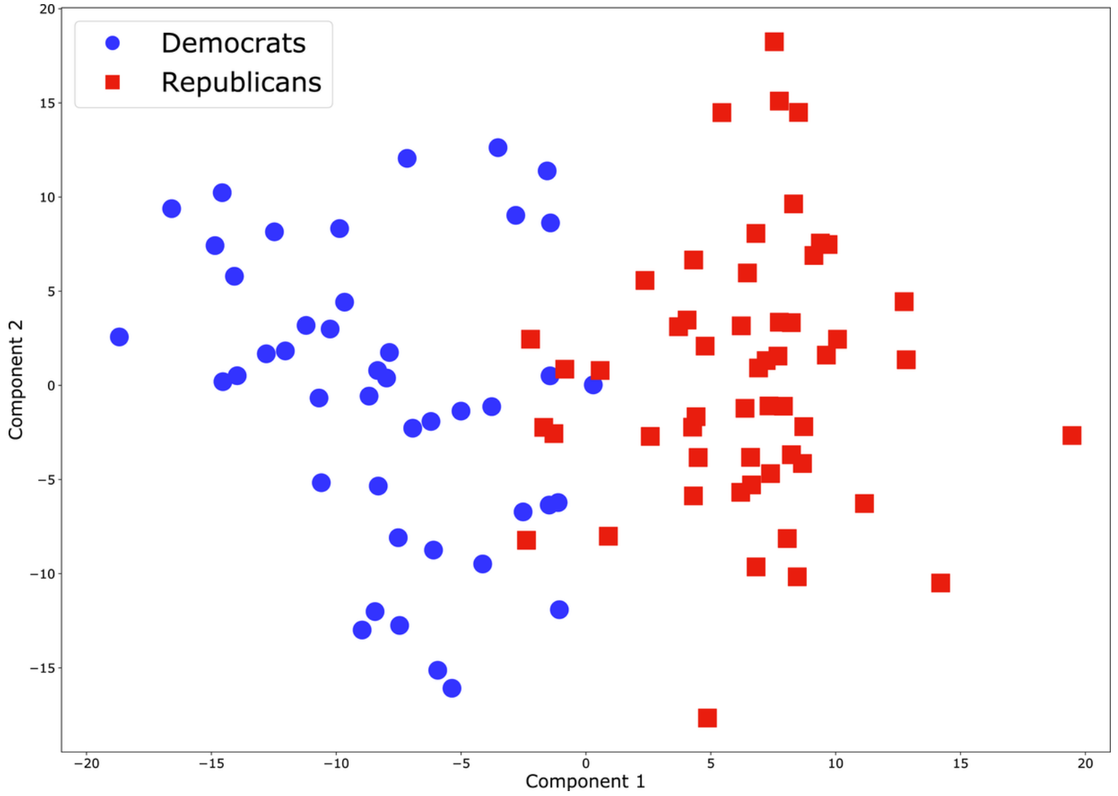
Metric	1921 - 2016			2007 -2016	
	Wordfish	Embeddings		Wordfish	Embeddings
Correlation	-0.003	0.93		0.015	0.999
Pairwise Accuracy	47.21%	85.83%		53.33%	91.11%

	1921–2016		2007–2016	
Metric	WordFish	Embeddings	WordFish	Embeddings
Correlation	−0.027	0.918	0.829	0.999
Pairwise accuracy	44.36%	85.66%	75.56%	88.89%

Figure 5. Ideological Placement of Senators



Replication Results: Senator Embedding



Author's Results: Senator Embedding

Table 4

Accuracy of Senator Ideological Placement

Gold Standard	Pearson Correlation	Updated Pearson Correlation	Pairwise Accuracy	Updated Pairwise Accuracy
DW-NOMINATE	0.876	0.87	80.81	80.96
Nokken-Poole	0.885	0.875	81.09	80.54
ACU 2016	0.852	0.841	75.38	74.58
ACU 2015	0.862	0.846	71.28	70.31
ACU Life	0.884	0.884	80.43	82.01
GovTrack	0.923	0.912	86.01	83.61

Autopsy

Backward compatibility

Analysis was originally implemented using Python 2.X and Gensim 3.X, thus we needed to implement code adjustments

Unavailability of corpus-specific pre-processing and training scripts

There are only sample codes for processing and training U.S. corpus, hence lack of visibility on corpus-specific pre-processing steps

The authors’ coding error in the pre-processing step

In the U.S. corpus, they defined a function that preprocesses whole line of the raw text data, rather than only the text column.

Autopsy

Lack of exact replication match

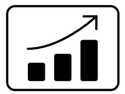
Not able to exactly replicate their findings due to differences in:

- Weights initializations
- Corpus and corpus pre-processing, especially in the British and Canadian Hansard data

Computational and storage cost

The pre-processed corpus has massive size, especially the US Congress corpus. Corpus pre-processing and model training are computationally costly. Both processes take 5-7 hours in total for each corpus.

Possible Extensions



Exploring party placement along political topics

Gauging the statistical significance of placement differences between political parties

Applying Kozlowski et al. (2019) word embedding geometry approach to project legislator position along specific issues

Analyzing political texts from other medium, i.e. media outlet ideologies, social media posts, etc.

Thank you!

Feel free to ask any questions!