FIGHTIN' WORDS

MILTON LIN

Contents

1. Dataset

We use a sample from the Congressional Record. This is a classical corpus used in many political science studies. Preprocessing has been done:

(1) The corpus was originally in plaintext format by [GST18] and prepared for NLP methods, specifically word2vec models, by [SR19]. Preprocessing includes removing non-alphabetic characters, converting text to lowercase, and removing words with a length of 2 or less.

(2) [SK22] further processed the plaintext R-data files into text (`txt`) and comma-separated values (`csv`) formats and subsampled the corpus for convenience.

The corpus now includes only Congressional sessions 111-114 (January 2009 - January 2017) and speeches by speakers with party labels "D" (Democratic) and "R" (Republican). Additionally, we consider a list of `politics_words`. These are predefined to be *freedom*, *justice*, *equality*, and *democracy*; partisan political issues like *abortion*, *immigration*, *welfare*, and *taxes*; as well as terms related to political parties, specifically *democrat* and *republican*.

Date: February 11, 2024.

2. Case study on Democratic and Republican speeches

We apply the method of Monroe et al. on logs odds ratio with Dirichlet prior, [MCQ08], for Democratic and Republican speeches.

Word	**Odds**	**Men Republican Count**	**Women Democrat Count**
women	69.6291	11281	16997
families	39.8979	16810	13278
children	37.5181	15868	12261
violence	35.6598	3076	4545
womens	35.1913	863	2700
her	35.1497	24944	16129
communities	32.6681	6898	6527
our	30.3342	203757	84229
african	29.6961	926	2208
for	29.5311	363359	141034

Table 1. Top 10 Words Favoring Women Democrats

Word	**Odds**	**Men Republican Count**	**Women Democrat Count**
spending	-36.5698	37546	3754
that	-35.3575	767275	209975
going	-32.1725	87089	16525
obamacare	-28.4968	12628	401
president	-28.2553	104206	22417
gentleman	-27.0023	38120	5941
taxes	-26.8178	18412	1662
you	-26.6062	147723	35042
government	-26.6017	62144	11953
trillion	-26.3713	16957	1442

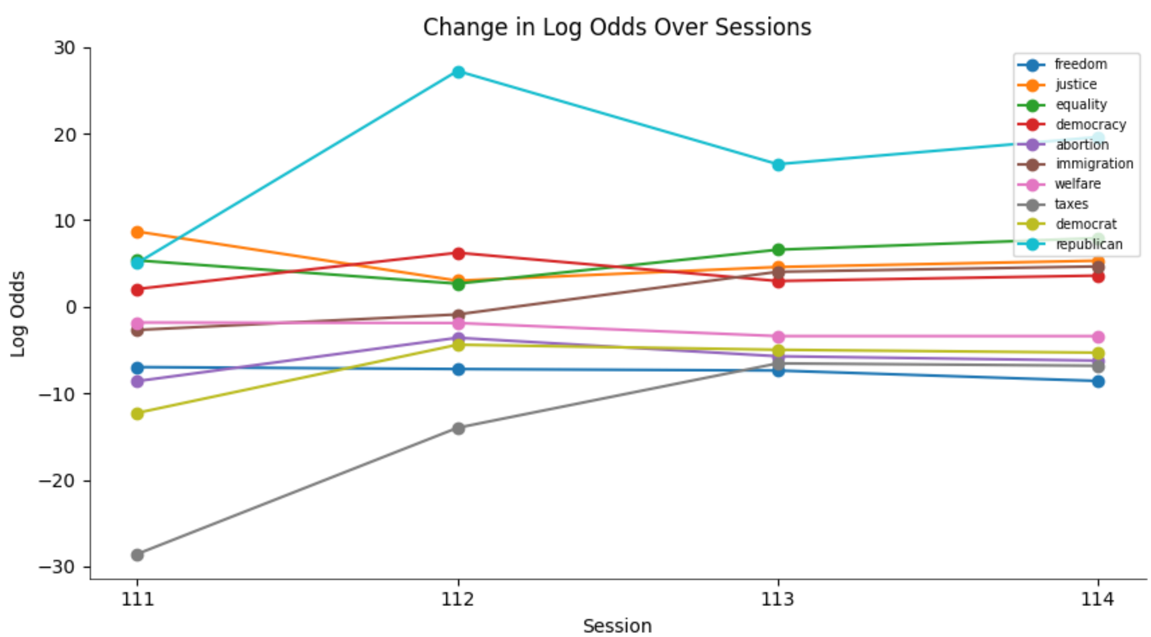
Table 2. Top 10 Words Favoring Men Republicans

2.1. **Time evolution.** We can apply moving average to see time evolution [MCQ08, p. 4.3]. Each row represents a word, and the columns represent different sessions. The last column is the maximum pairwise absolute change.

Word	111	112	113	114	Max Change
freedom	-6.97042	-7.19641	-7.36784	-8.57041	1.59999
justice	8.67716	3.00529	4.58445	5.31562	5.67187
equality	5.35082	2.65929	6.58165	7.91783	5.25854
democracy	2.05033	6.2351	2.96912	3.59459	4.18476
abortion	-8.59089	-3.58323	-5.71722	-6.2239	5.00766
immigration	-2.68338	-0.89142	4.03172	4.65159	7.33497
welfare	-1.82258	-1.88037	-3.39009	-3.39299	1.57041
taxes	-28.5482	-13.9838	-6.54389	-6.81583	22.0043
democrat	-12.2578	-4.38769	-4.96859	-5.30596	7.87007
republican	5.05606	27.224	16.4564	19.5824	22.1679

TABLE 3. Changes over time in log odds with prior and maximum absolute changes

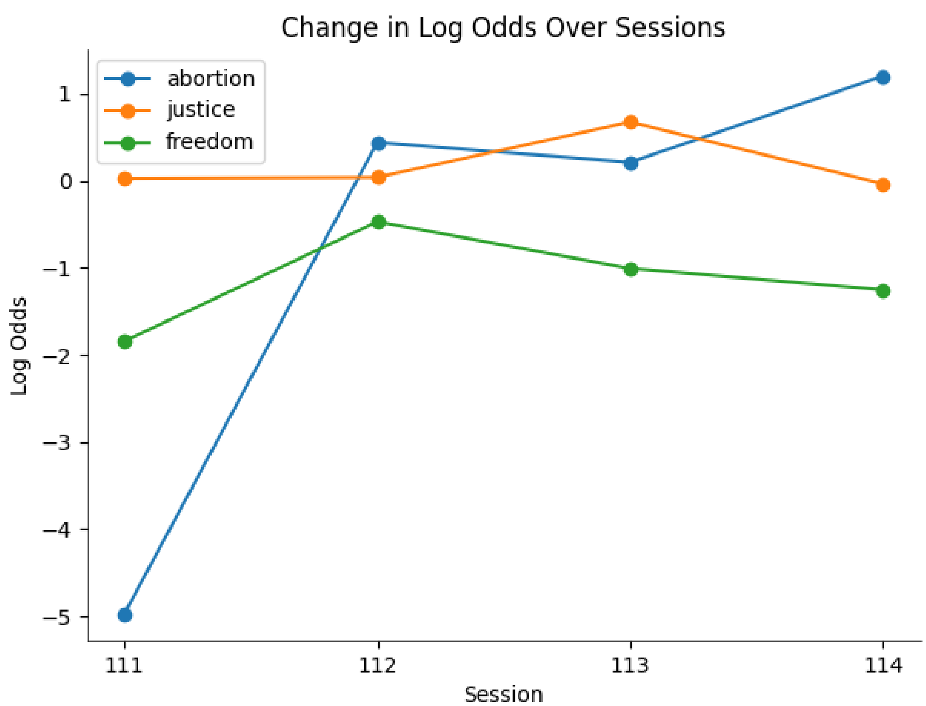
Wse see that the meaning of the words "taxes" and "republican" has changed significantly over time. And the word "immigration" has turned from a more Republican to a more Democratic one.



2.2. **Latent dirichlet allocation on healthcare.** Next we apply the method of LDA **??**, to understand the effect of logs odd when we restrict within a topic (the topic of choice is **healthcare**) and not.

Word	**111**	**112**	**113**	**114**
abortion	-4.97293	0.439798	0.212665	1.20238
justice	0.0280993	0.0398734	0.674044	-0.0346049
freedom	-1.84361	-0.470894	-1.00766	-1.25194

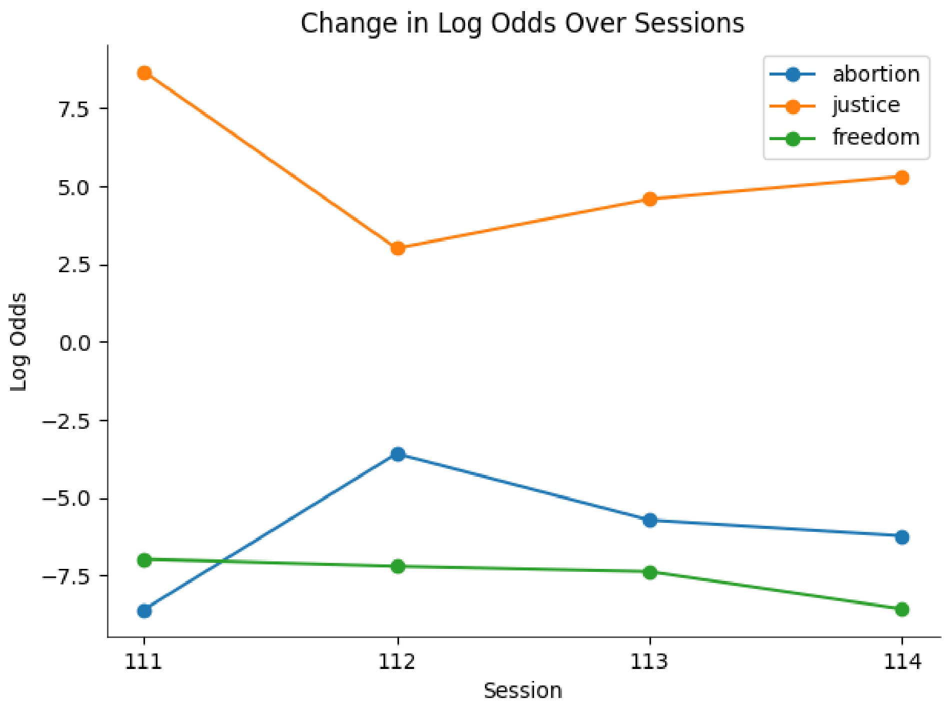
TABLE 4. Change in Log Odds in **healthcare** topic



In comparison with out topic, we have

Word	**111**	**112**	**113**	**114**
abortion	-4.97293	0.439798	0.212665	1.20238
justice	0.0280993	0.0398734	0.674044	-0.0346049
freedom	-1.84361	-0.470894	-1.00766	-1.25194

TABLE 5. Change in Log Odds Over All Documents



- *Abortion:* Within the topic of healthcare, the word "abortion" has positive log odds, indicating its higher frequency in Democratic speeches. However, considering all contexts, it is used more frequently in Republican speeches. This discrepancy suggests that the word's usage varies significantly between the two parties, especially within healthcare discussions.
- *Scale of Log Odds:* Within the healthcare context, we observed that the log odds are at a slightly smaller scale compared to the analysis without a specific topic. This observation is consistent with the idea that a more focused context, such as healthcare, results in a narrower range of word usage.
- *Justice and Freedom:* The usage of the words "justice" and "freedom" does not show a clear preference by either political party within the healthcare topic, particularly for the former. However, when analyzing log odds without considering a specific context, the usage patterns may differ significantly. [1]

These observations highlight the importance of context when analyzing word usage and log odds, as it can significantly impact the interpretation of linguistic patterns within political speeches.

3. Word embeddings

The text has already been preprocessed, [SR19] for the purpose of word2vec. Let us briefly described the parameters we used `Word2Vec` model.

[1] Perhaps this suggests that these words may have different connotations and usage patterns when discussed in healthcare-related contexts compared to general discussions.

- `workers`: the number of threads used. [2]
- `seed`: the word2vec begins by initializing random vector for each word.

We now train a Word2Vec model on the Republican and Democratic text data, respectively. We consider `query`="taxes".

Word	Similarity
tax	0.7361
taxing	0.6145
surtax	0.6051
taxation	0.5506
revenue	0.5407
raise	0.5323
earners	0.5318
inequality	0.5310
taxed	0.5204
raising	0.5176

TABLE 6. Republican Near Neighbors to Taxes

Word	Similarity
tax	0.7226
revenues	0.6387
revenue	0.6323
taxed	0.6277
taxing	0.6013
taxation	0.5911
pay	0.5772
excise	0.5712
paying	0.5687
fica	0.5321

TABLE 7. Democrat Near Neighbors to Taxes

Analysis:

- The top 10 words have a lot of repeating/morphological variants. It is perhaps better to take more words from each list or apply methods such as lemmatization to provide a more diverse set.
- *Republicans* words like "surtax", "raise", "earners", and "inequality" appear. This might indicate a focus on the implications of taxes on earnings and economic outcomes.
- *Democrat* words. Words like "revenues", "pay", "excise", and "fica" suggest taxes in the context of public services and social issues.

[2]parallel processing introduces non-determinism in the order in which words are processed, which can affect the final embeddings slightly.

3.1. **Aligning word spaces.** The embeddings from Democratic and Republican sources (`d_embs` and `r_embs`, respectively) are aligned using the `align_matrices` function. For each word in a list of politically relevant terms (`politics_words`), we compute the cosine similarity between the embeddings of that word in the Democratic and Republican aligned spaces.

Political Word	**Similarity Score**
Freedom	0.8327
Justice	0.8293
Democracy	0.8122
Immigration	0.7619
Equality	0.7373
Taxes	0.7315
Abortion	0.7072
Welfare	0.6584
Democrat	0.6073
Republican	0.6020

Table 8. Similarity scores of political words between Democrat and Republican corpora

- *High similarity* Words like "freedom," "justice," and "democracy" have high similarity scores (over 0.8), suggesting a shared value system.
- *low similarity* "Welfare," "democrat," and "republican" exhibit lower similarity scores (below 0.66). This might reflect divergent viewpoints or policies

3.2. **Average cosine distance across different congressional sessions.**

Congressional Session	**Average Cosine Similarity**
111	0.685958206653595
112	0.6810095906257629
113	0.7146731615066528
114	0.6955482363700867

Table 9. Average cosine similarity of key political words across congressional sessions

Analysis:

- *The average cosine similarity of key political words across congressional sessions shows minor fluctuations.* The similarity scores start at 0.6859 in session 111 and then to 0.6955 in session 114.
- *The increase in average cosine similarity in session 113 suggests a closer alignment* in the use of specific political terms between the two parties during that session.

Let us comment on the many limitations of this approach:

- *Selection of Political Words:* The analysis is limited by the predefined set of political words, which does not capture full political discourse. Additionally, the significance and connotations of these words may change over time!
- *Contextual nuances and ambiguity:* while the embeddings here are indeed domain specific, political language is often context dependent. For instance, thought the words are used differently, it does not show the irony, emotion and sarcasm within the speakers.
- *A more granular time-based analysis:* could provide deeper insights into polarization trend, since political events often is quite dynamic in short periods.

References

[GST18] Gentzkow, Matthew, Shapiro, Jesse M, and Taddy, Matt. "Congressional record for the 43rd-114th congresses: Parsed speeches and phrase counts". In: *URL: https://data. stanford. edu/congress text*. 2018 (cit. on p. 1).

[MCQ08] Monroe, Burt L, Colaresi, Michael P, and Quinn, Kevin M. "Fightin'words: Lexical feature selection and evaluation for identifying the content of political conflict". In: *Political Analysis* 16.4 (2008), pp. 372–403 (cit. on pp. 2, 3).

[SK22] Stewart, Ian and Keith, Katherine. "Democratizing Machine Learning for Interdisciplinary Scholars: Report on Organizing the NLP+ CSS Online Tutorial Series". In: (2022) (cit. on p. 1).

[SR19] Spirling, Arthur and Rodriguez, Pedro L. "What works, what doesnt, and how to tell the difference for applied research". In: (2019) (cit. on pp. 1, 5).