Party Identification Through Bill Summaries

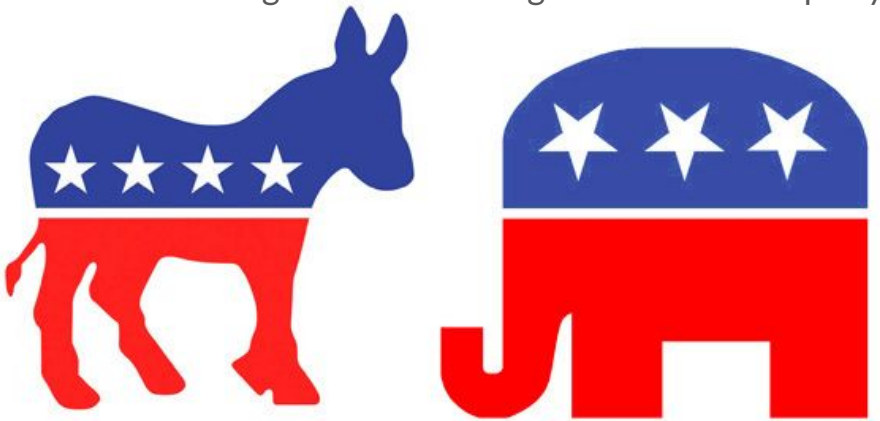
Mitchell Joseph & Maia Austin

Research Question

Can we build a bill text classifier, which, when given a bill summary from the US House of Representatives, can correctly classify the bill as having been sponsored by either a Republican or a Democrat.

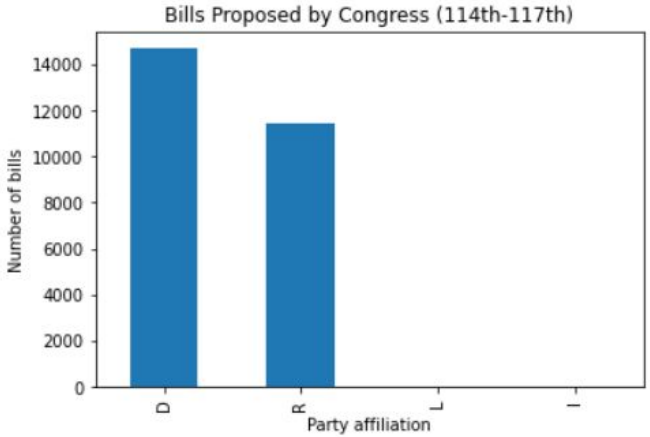
Motivation

- Political ideology has become increasingly polarized
- Can we capture the biases within each party's ideology
- Can we get a better understanding of the motivating issues that drive party politics



Data

- We built a custom webscaper that collected bill summary data from the 114th-117th Congress[1]
- Congress' website
- We were able to obtain over 26,000 samples
- Everything stored in a SQL database[2]



1. 2015-2021
2. https://github.com/Mitch-ml/CS254/tree/main/data

Working With Text

- Algorithms expect numerical features instead of raw text
- There are many ways to get around this but we used a Bag-of-Words approach

Bag-of-Words

- Assign an integer id to each word within your training data
- Count the number of occurrences of each word
- The number of features will equal the number of distinct words in the corpus

	the	red	dog	cat	eats	food
1. the red dog →	1	1	1	0	0	0
2. cat eats dog →	0	0	1	1	1	0
3. dog eats food →	0	0	1	0	1	1
4. red cat eats →	0	1	0	1	1	0

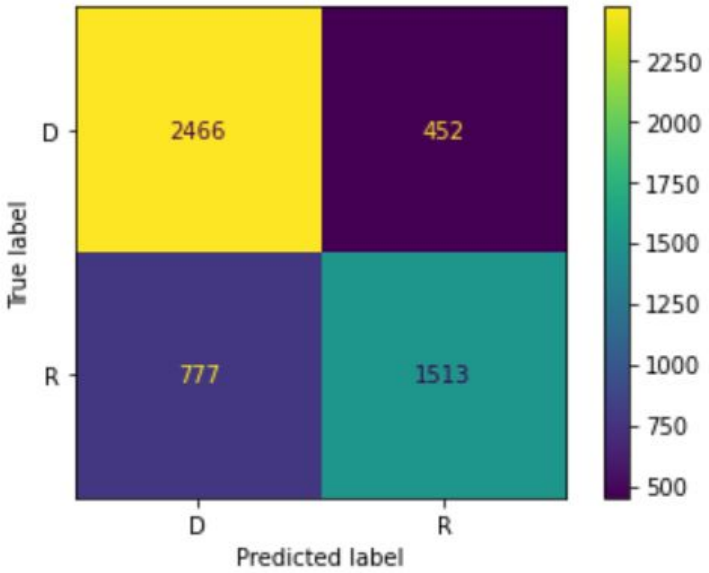
Term Frequency times Inverse Document Frequency

- **Problem:** Longer documents will have a higher average count values than shorter documents even when they are talking about the same issue
- **Solution:** Divide the number of occurrences of each word in a document by the total number of words in the document (*Term Frequencies*)
- We also downscale weights for words that occur in many documents in the corpus, which are therefore less informative
- This downscaling is called TF-IDF

Methodology

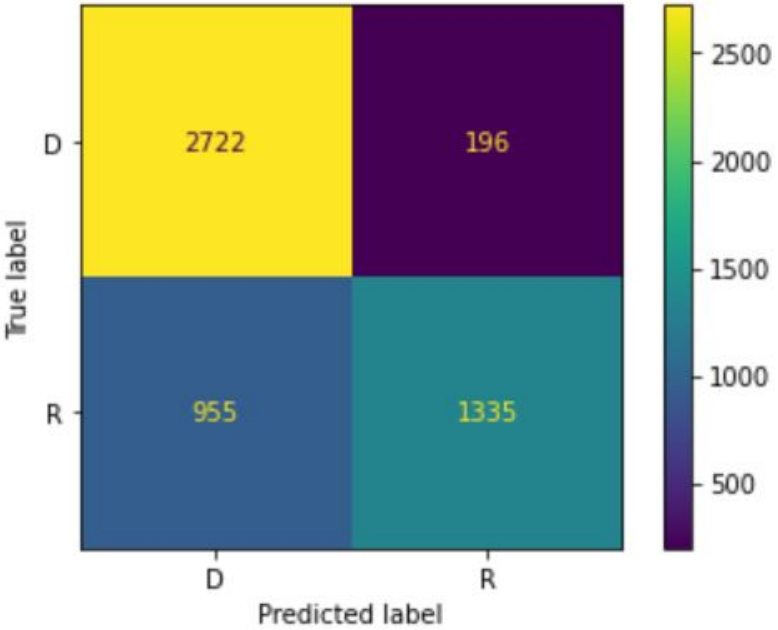
- 80/20 split
- Bag-of-Words approach
- Vectorized our words using TF-IDF
- Compared several models including:
    - Logistic Regression
    - Multinomial Naive Bayes
    - Support Vector Classifier
    - Support Vector Machine
    - Random Forest

Logistic Regression



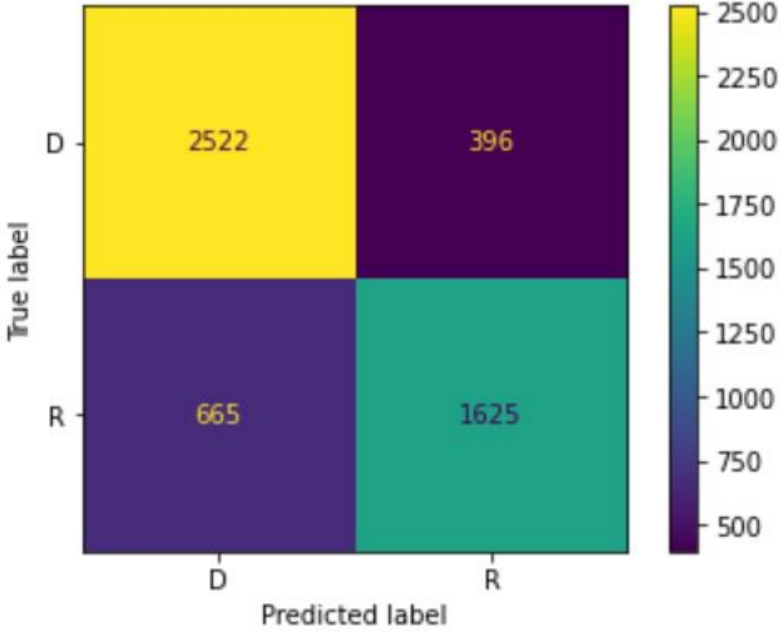
	precision	recall	f1-score	support
D	0.76	0.85	0.80	2918
R	0.77	0.66	0.71	2290
accuracy			0.76	5208
macro avg	0.77	0.75	0.76	5208
weighted avg	0.76	0.76	0.76	5208

Multinomial Naive Bayes



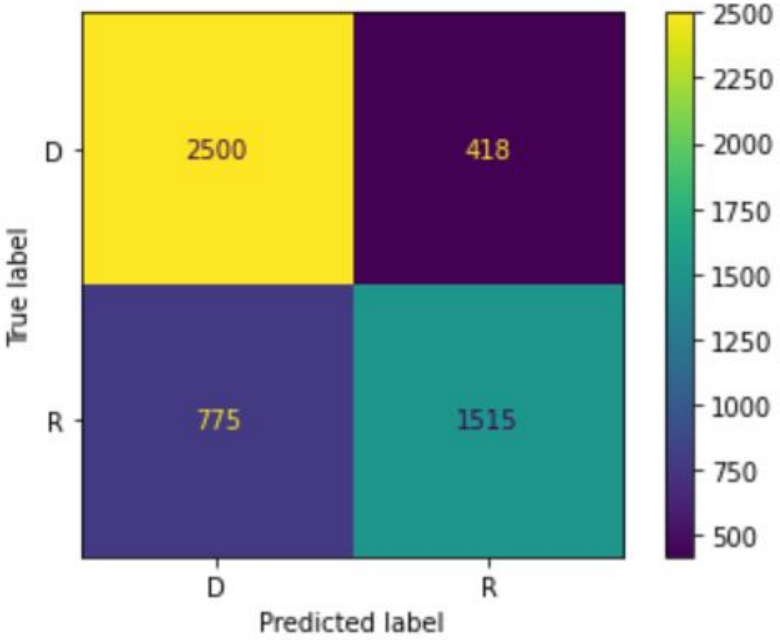
	precision	recall	f1-score	support
D	0.74	0.93	0.83	2918
R	0.87	0.58	0.70	2290
accuracy			0.78	5208
macro avg	0.81	0.76	0.76	5208
weighted avg	0.80	0.78	0.77	5208

SVC



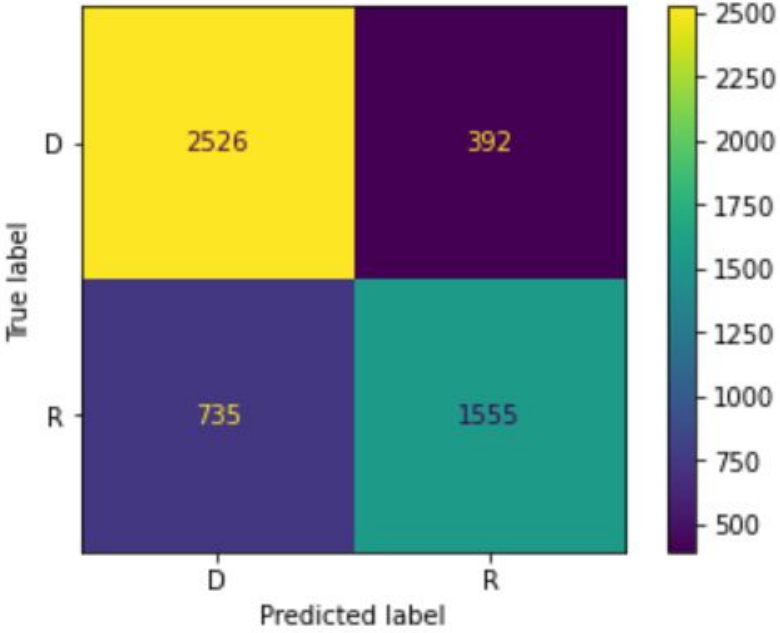
	precision	recall	f1-score	support
D	0.79	0.86	0.83	2918
R	0.80	0.71	0.75	2290
accuracy			0.80	5208
macro avg	0.80	0.79	0.79	5208
weighted avg	0.80	0.80	0.79	5208

Linear SVM



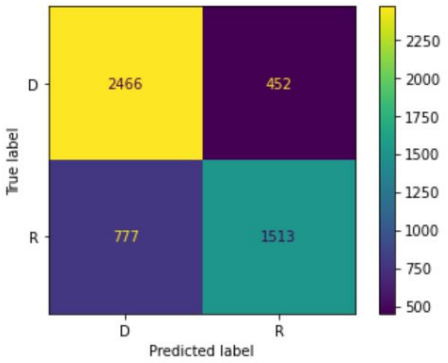
	precision	recall	f1-score	support
D	0.76	0.86	0.81	2918
R	0.78	0.66	0.72	2290
accuracy			0.77	5208
macro avg	0.77	0.76	0.76	5208
weighted avg	0.77	0.77	0.77	5208

Random Forest

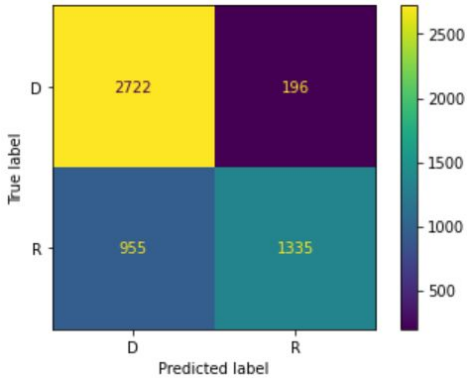


	precision	recall	f1-score	support
D	0.77	0.87	0.82	2918
R	0.80	0.68	0.73	2290
accuracy			0.78	5208
macro avg	0.79	0.77	0.78	5208
weighted avg	0.79	0.78	0.78	5208

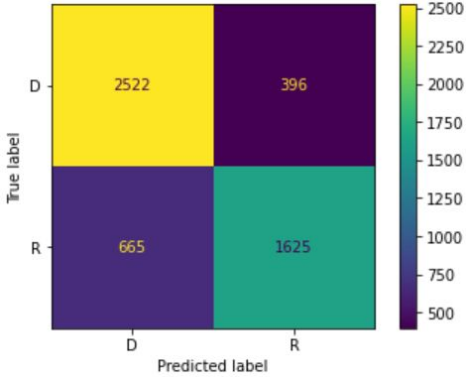
Logistic Regression



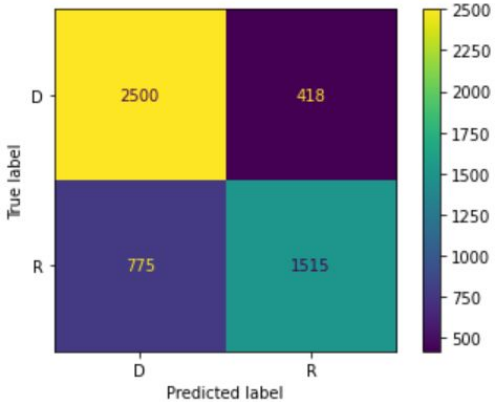
Multinomial NB



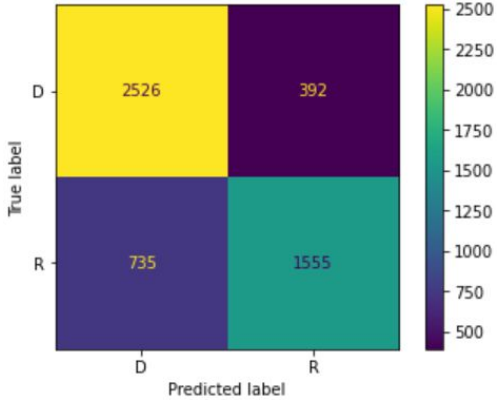
SVC



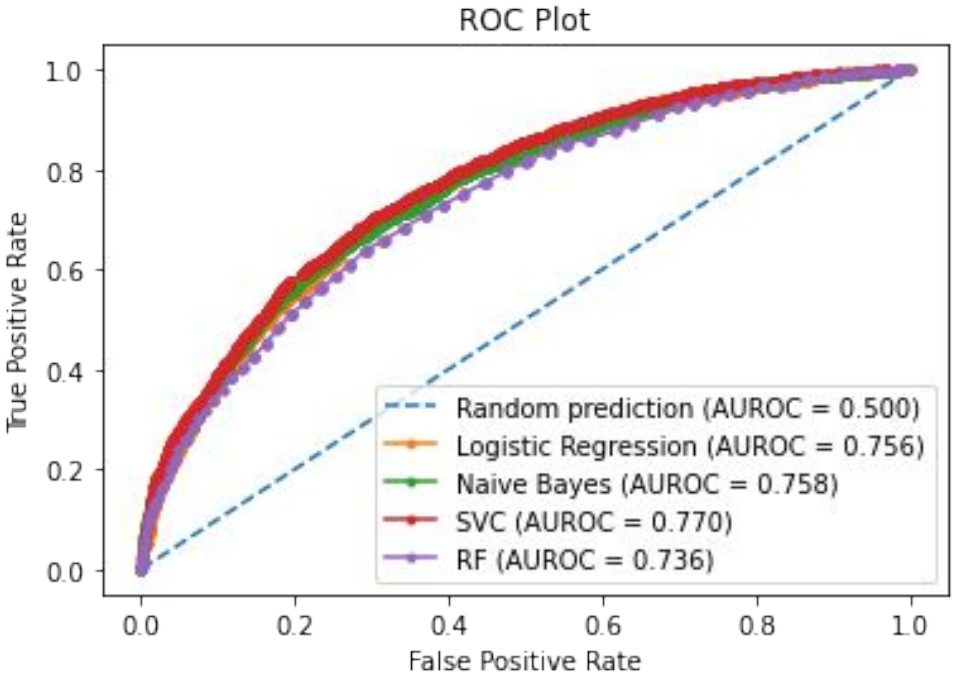
Linear SVM



Random Forest



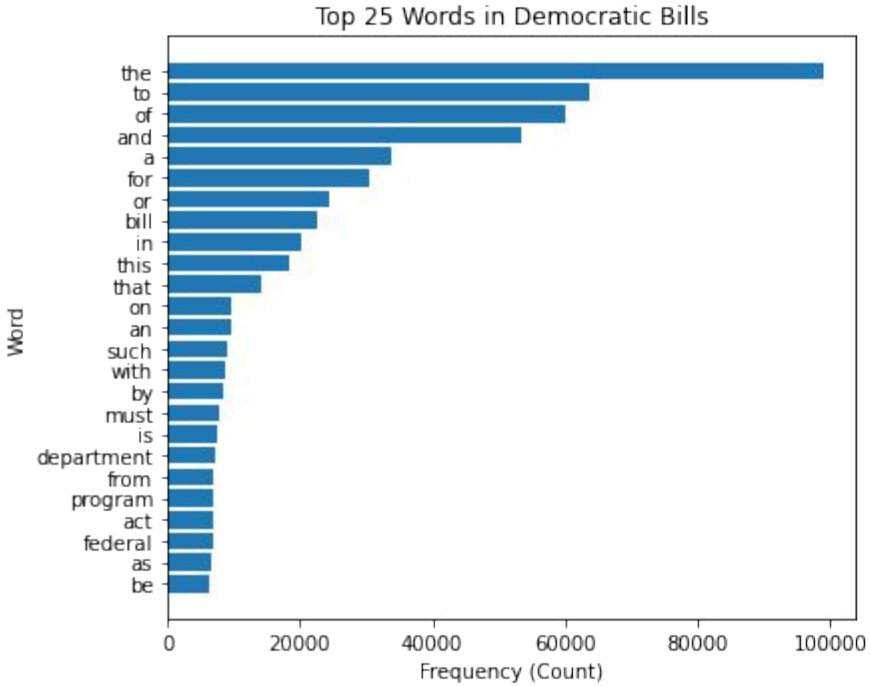
ROC Curves

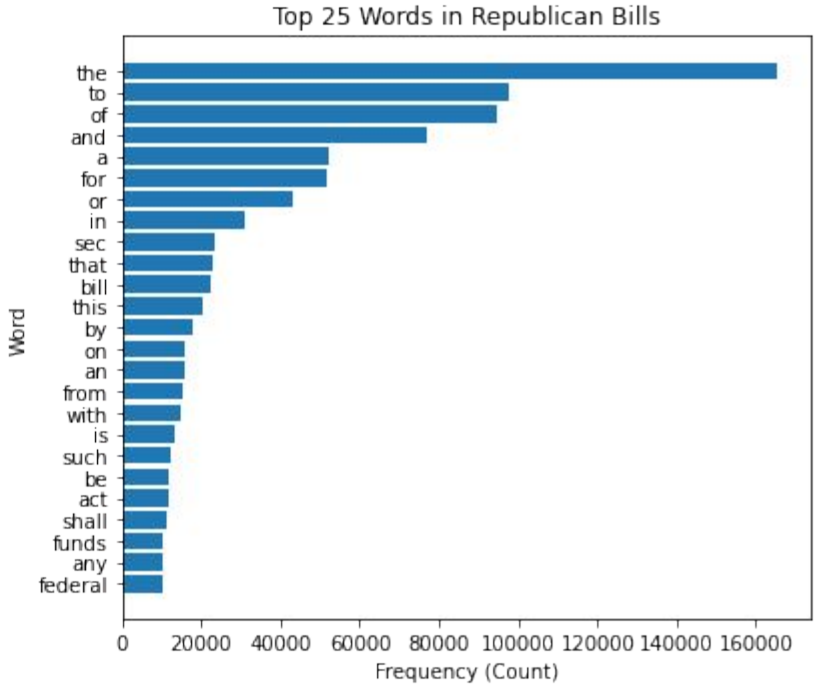

ROC Plot
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
Random prediction (AUROC = 0.500)
Logistic Regression (AUROC = 0.756)
Naive Bayes (AUROC = 0.758)
SVC (AUROC = 0.770)
RF (AUROC = 0.736)

Exploratory Data Analysis

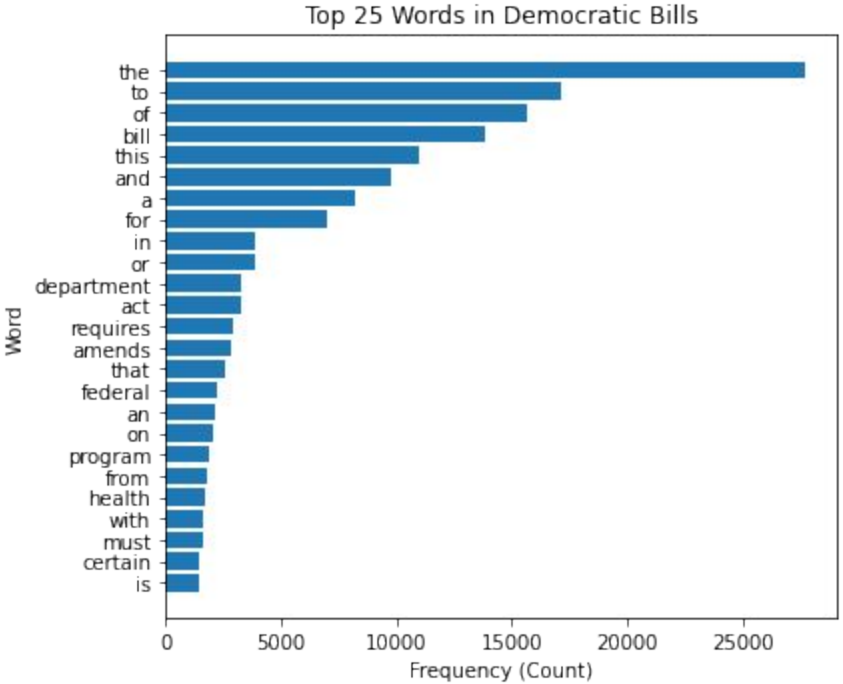
- Word frequencies
- HuggingFace 
- Shifterator Plots

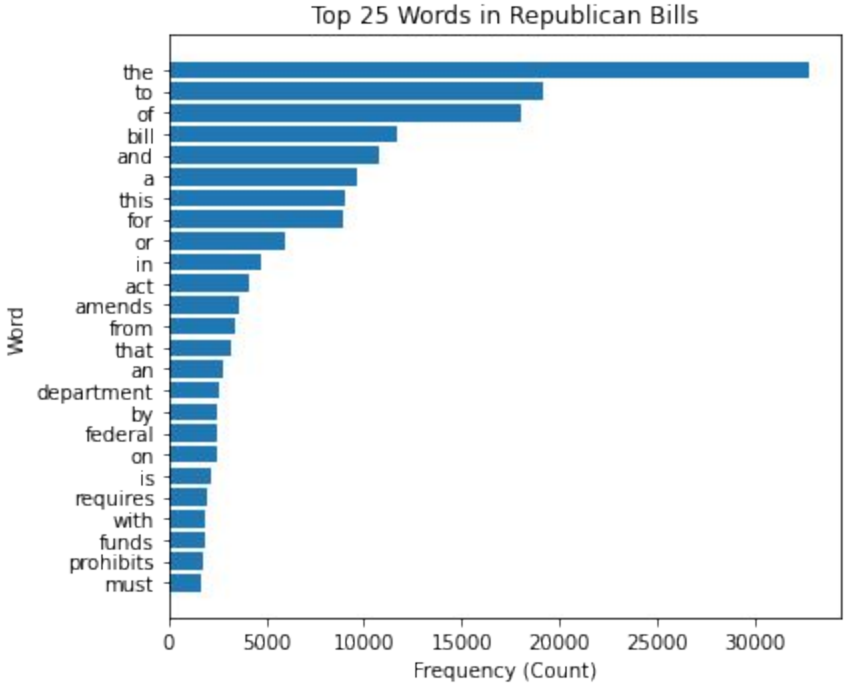
Top 25 Words (no stop words removed)



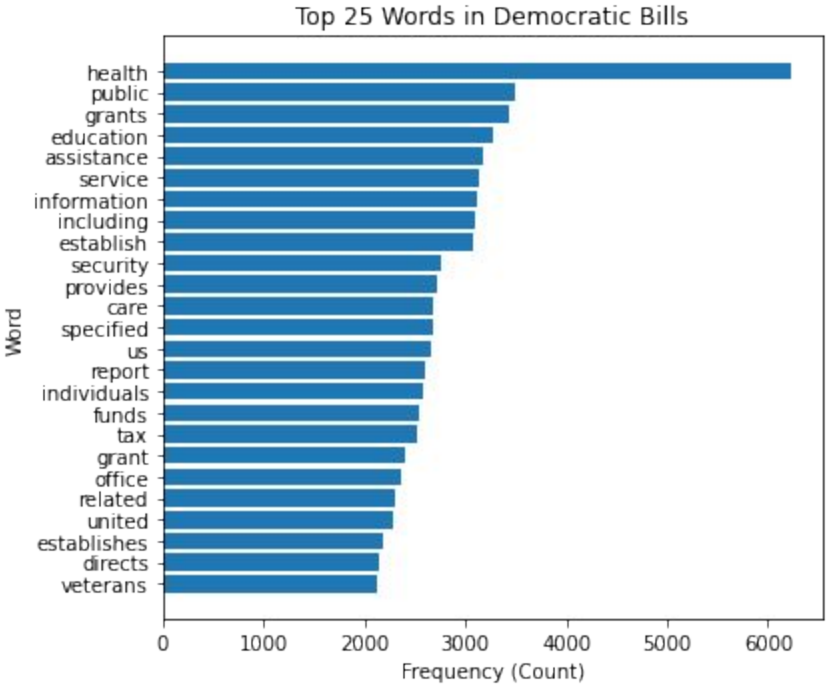


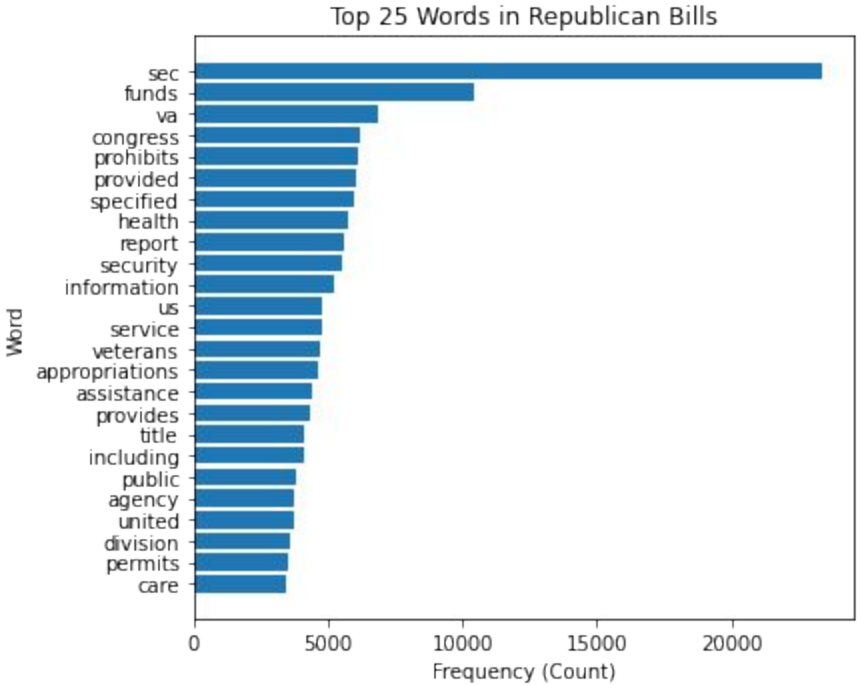
Top 25 Words HF (no stop words removed)



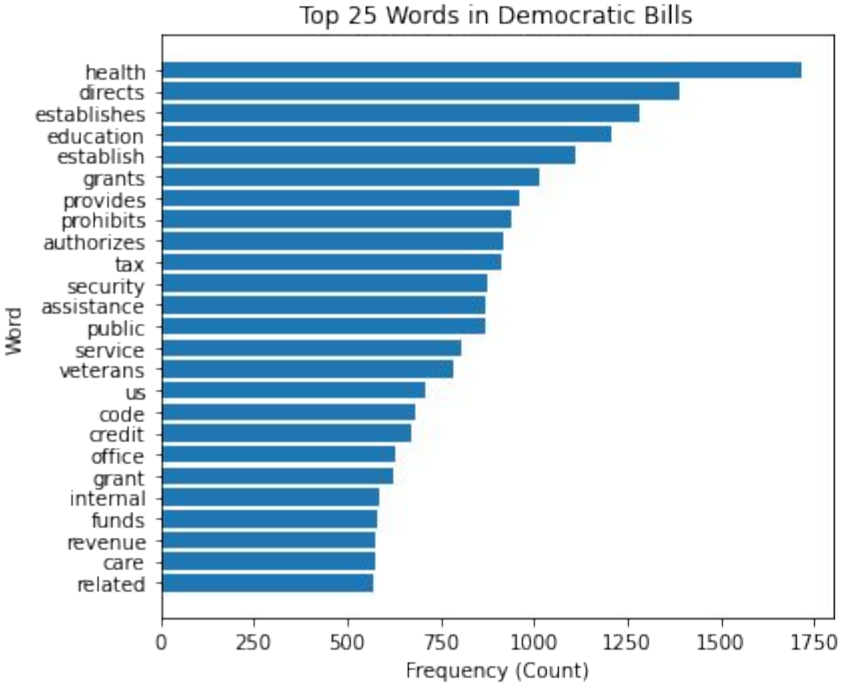


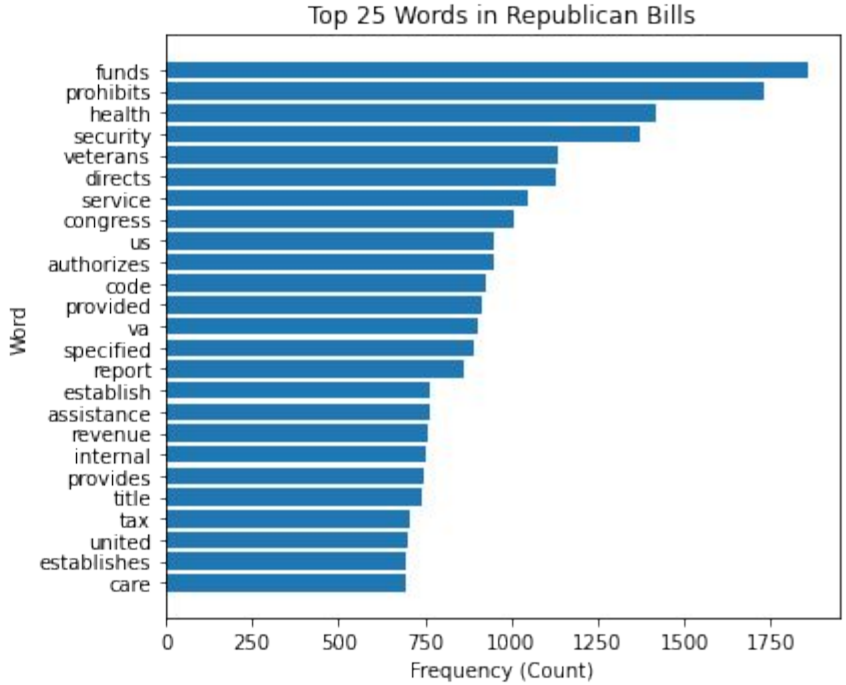
Top 25 Words (stop words removed)





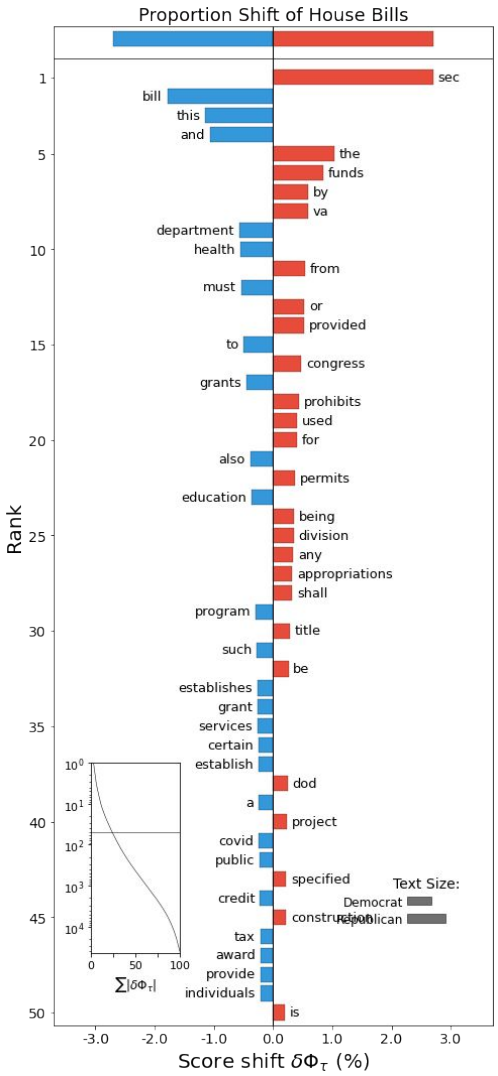
Top 25 Words HF (stop words removed)

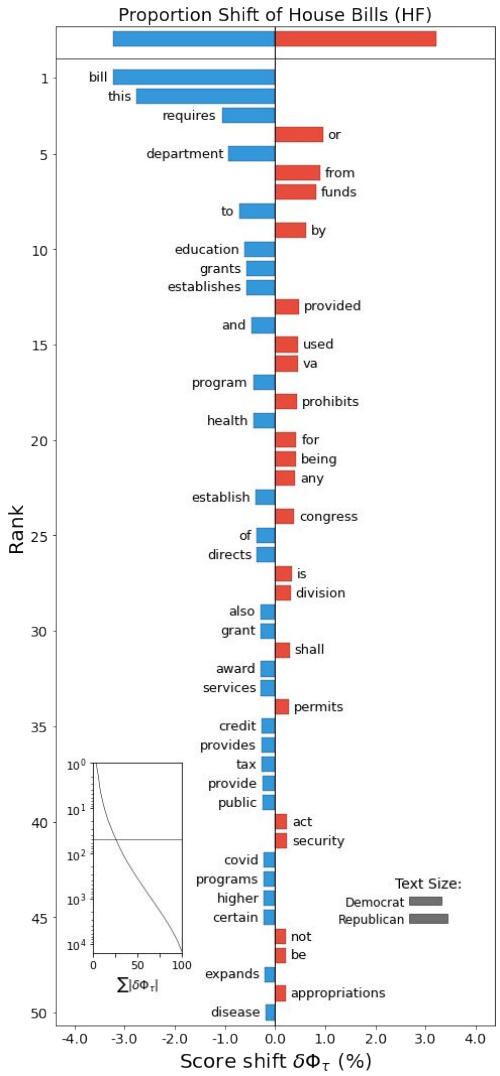




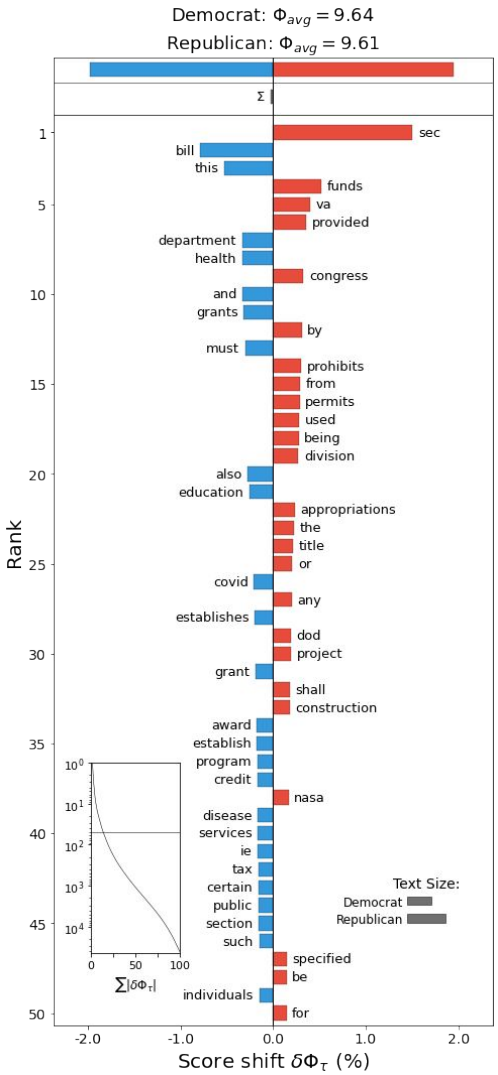
Shifterator Plots

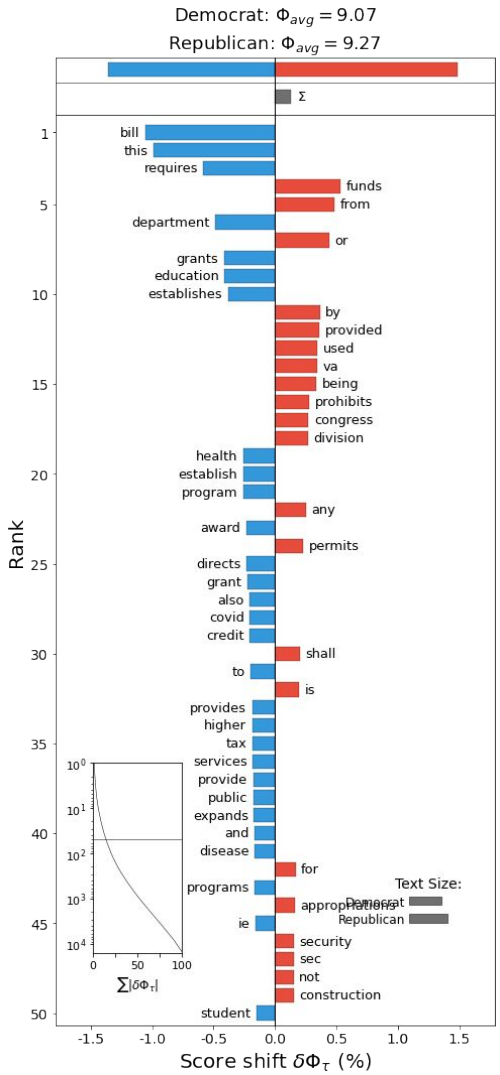
- **Proportional Shifts:** Calculates the difference in relative frequencies of words.
  - We can now visualize if a word is relatively more common in one text vs the other
- **Shannon Entropy Shifts:** A way to identify more “surprising” words and how they vary between two texts.
  - We’re given an entropy score - think of this as how unpredictable the text is
- There will be a cumulative contribution plot in the lower left
  - This shows how much of the overall difference is explained by the top contributing words
- The other plot will show the relative text size of each corpus by the number of tokens


Proportion Shift of House Bills
Rank
Score shift $\delta\Phi_\tau$ (%)
sec
bill
this
and
the
funds
by
va
department
health
from
must
or
provided
to
congress
grants
prohibits
used
for
also
permits
education
being
division
any
appropriations
shall
program
title
such
be
establishes
grant
services
certain
establish
dod
a
project
covid
public
specified
credit
construction
tax
award
provide
individuals
is
Text Size:
Democrat
Republican
$\sum|\delta\Phi_\tau|$


Proportion Shift of House Bills (HF)
Rank
Score shift $\delta\Phi_\tau$ (%)
bill
this
requires
or
department
from
funds
to
by
education
grants
establishes
provided
and
used
va
program
prohibits
health
for
being
any
establish
congress
of
directs
is
division
also
grant
shall
award
services
permits
credit
provides
tax
provide
public
act
security
covid
programs
higher
certain
not
be
expands
appropriations
disease
Text Size:
Democrat
Republican
$\sum|\delta\Phi_\tau|$

Full text summary





HF summarized data

Conclusion

- SVC had the best classification
  - 80% accuracy
  - 79% F1-score
  - 77% AUROC
- Using a transformer to summarize the data removed too much information to be useful in classification

Questions

Bibliography

Boche, A., Lewis, J. B., Rudkin, A., & Sonnet, L. (2018). The new Voteview.com: Preserving and continuing Keith POOLE'S infrastructure for scholars, students and observers of Congress. Public Choice, 176(1-2), 17–32. https://doi.org/10.1007/s11127-018-0546-0

Gallagher, R. J., Frank, M. R., Mitchell, Lewis, Schwartz, A. J., Reagan, A. J., Danforth, C. M., Dodds, P. S. (2021). Generalized Word Shift Graphs: A Method for Visualizing and Explaining Pairwise Comparisons Between Texts. EPJ Data Science, 10(4).

Grimmer, J. (2010). A Bayesian Hierarchical topic model for Political Texts: Measuring Expressed agendas in SENATE press releases. Political Analysis, 18(1), 1–35. https://doi.org/10.1093/pan/mpp034

Laver, M., Benoit, K., & Garry, K. (2003). Extracting Policy Positions from Political Texts Using Words as Data. *American Political Science Review, 97*(2), 311-331. doi:10.1017/S0003055403000698

Lin, Y. R., Margolin, D., & Lazer, D. (2015). Uncovering social semantics from textual traces: A theory‐driven approach and evidence from public statements of U.S. members of congress. Journal of the Association for Information Science and Technology, 67(9), 2072–2089. https://doi.org/10.1002/asi.23540

Sim, Y., Acree, B., Gross, J., Smith, N. (2013). Measuring Ideological Proportions in Political Speeches. Empirical Methods in Natural Language Processing (EMNLP).

Y, Bei., Kaufmann, S., Diermeier, D. (2008) Classifying Party Affiliation from Political Speech, Journal of Information Technology & Politics, 5:1, 33-48, http://dx.doi.org/10.1080/19331680802149608