Tweeting Across the Aisle: Analyzing Political Framing and Polarization in U.S. Senator Tweets Using Word Embeddings

Ella Happel
University of Illinois Urbana-Champaign / Champaign, IL
ehappel2@illinois.edu

Abstract

This paper explores the framing of political ideas and issues in U.S. politics by the Democratic and Republican parties through the analysis of partisan word embeddings trained on a dataset of 99,693 tweets. It finds that there is substantial framing differences in relevant ideas and issues between parties, with an average Jaccard similarity score of 0.0367 and an average cosine similarity score of 0.1589 for chosen words. It explores the quantitative similarities for each keyword, and qualitatively evaluates the different nearest neighbors of the keywords to form a greater picture of each parties' lexicon. Keywords *abortion*, *capitalism*, *climate change*, *environment*, *gun*, *immigrant*, *immigration*, *inflation*, *justice*, *medicare*, *police*, *racism*, *security*, *tax*, *terrorism*, and *voting right* show stark partisan divergence (Jaccard and cosine similarity scores of 0), while keywords *education, economy, equality, freedom, military, patriotism, right, truth*, and *veteran* reported the highest similarity between parties.

1 Introduction

As we move into the digital era, the great influential speeches have moved digital as well. Politicians have taken to sites like X (formerly Twitter) to share their beliefs and outlooks on the world, with millions following along. This online data is a goldmine of information on how political figures, and the parties they belong to, speak about political issues and ideas. In this paper we will seek to uncover the differences in framing of those ideas by U.S. Democrat and Republican senators, in hopes to understand each parties' true positioning and use of important words.

The training data is sourced from a dataset senator-tweets, containing 99,693 tweets by 99 US senators from 1/20/2021 to 12/30/2021. In the following sections, we will preprocess this data by tokenizing, combining phrases, and lemmatizing. Then we will choose a set of 32 keywords to explore for the duration of the paper, and create Word2Vec Skip-gram word embeddings of each word, for each party (Democrat or Republican). We will plot the 10 nearest neighbors, and qualitatively explore the differences in nearest neighbors between parties. Then, we will apply Jaccard similarity and cosine similarity measures to quantitatively identify which keyword embeddings were similar or different between parties, noting lexical or semantic differences based on the measure.

2 Related Work

Sentence level embeddings of this dataset, with analysis of liberal to conservative views of topics amongst senators is explored in (Chen et al., 2024). This paper expands upon this work, focusing instead on word-level embeddings and neighborhoods.

3 Data Exploration

You can access our code pipeline here, and the senator-tweets dataset here.

Our data comes from the senator-tweets dataset, created by Newhauser, from the Hugging Face dataset library (Lhoest et al., 2021). Each entry stores the date, tweet id, username, text, party, and a tweet embedding. For the purposes of our analysis, we will use the username, text, and party variables.

We chose 32 keywords to explore in this paper, comparing the different usages of each by each party. We aimed to choose important and relevant issues in American politics (like immigration and healthcare), alongside more abstract concepts of US politics (like freedom and justice). We selected the following keywords:

- abortion	- border	- capitalism
- climate_change	- crime	- democracy
- economy	- education	- environment
- equality	- freedom	- gas
- gun	- healthcare	- immigrant
- immigration	- inflation	- job
- justice	- medicare	- military
- patriotism	- police	- racism
- right	- security	- social_security
- tax	- terrorism	- truth
- veteran	- voting_right	

To clean and preprocess our Twitter data, we tokenized the tweets with NLTK and combined phrases with gensim. Then, we lemmatized with NLTK. This resulted in two lists of lists of strings–each list containing a list of sentences which contains a list of word tokens. We also lemmatized the keywords in the keyword list.

4 Representations

To create our analyses, we built Skip-gram Word2Vec embeddings of each keyword, using the preprocessed lists of data. Skip-gram is utilized here for its effectiveness of capturing semantic relationships. The seed is set to 42 and workers to 1 to help with future replication of our pipeline. Vector size is 100 and window size is 5.

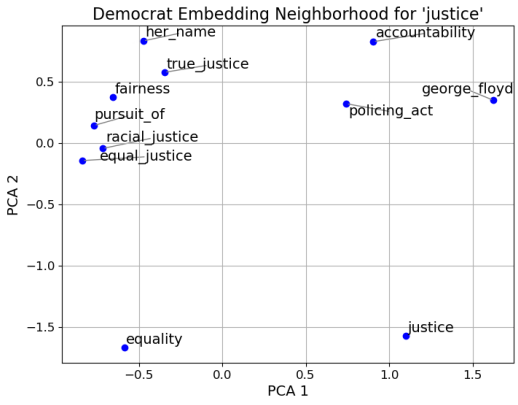

Figure 1: Democrat Embedding Neighborhood for "Justice"

Then, we found the 10 nearest neighbors for each word in each party, and plotted them on a graph using principal component analysis, reducing to 2-dimensions. We also saved the 10 nearest neighbors to a csv file for easier exploration. Included in this paper are 3 visualizations from this analysis: a graph for the Democratic and Republican nearest

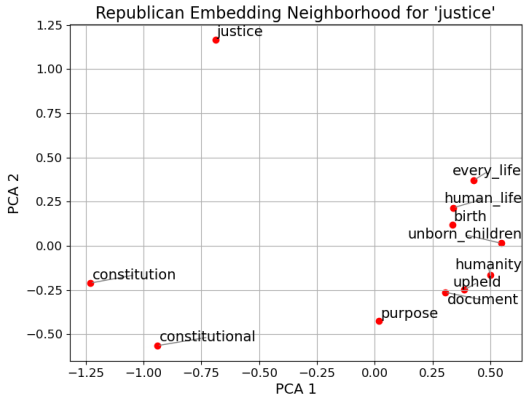

Figure 2: Republican Embedding Neighborhood for "Justice"

neighbors of *justice*, and a table of their nearest neighbors with their similarity (Figure 1, Figure 2, Table 1). The Democratic embeddings for *justice* center around themes of equality, police, and racial violence, while republicans referenced the constitution and abortion.

Generally, some example qualitative findings by keyword from this analysis are:

Freedom: Republicans referenced relevant populations like Cuban People and Hong Kongers, alongside communism and religious liberty. Otherwise similar connotations of fundamental human right and liberty.

Security: Republicans had highest word-similarity in national security and sovereignty, while democrats had highest word similarity in stability and cyber.

Immigrant: Democrats had associations with families, workers, and diversity, while Republicans had associations with crossing and illegal immigration.

Crime: Democrats referenced anti-asian hate and "military sexual", while Republicans referenced human trafficking. There is shared associations of sexual violence.

Climate Change: Democrats referenced hashtags #climatecrisis and #climatechange alongside climate effects, while Republicans had associations of energy sources and policy, fossil fuels, oil production, and spending.

5 Quantitative Similarity Comparisons

To compare the overlap between democrat and republican embeddings quantitatively, we created plots and a table of their cosine and Jaccard similar-

Democrat Nearest Neighbors	Similarity	Republican Nearest Neighbors	Similarity
equality	0.754	birth	0.745
policing_act	0.696	document	0.744
true_justice	0.671	constitution	0.735
racial_justice	0.665	upheld	0.734
fairness	0.659	human_life	0.723
accountability	0.659	constitutional	0.722
george_floyd	0.657	purpose	0.722
equal_justice	0.656	humanity	0.719
pursuit_of	0.646	unborn_children	0.706
her_name	0.634	every_life	0.701

Table 1: Nearest Neighbors for *justice* in Democrat and Republican Embeddings

ity (figure 3, figure 4, table 3). There are obvious similarities between the plots, with the same keywords having higher and lower similarity scores, however some keywords performed proportionately better or worse on one or the other. The cosine similarity output also had across-the-board much higher similarity scores, meaning Democrats and Republicans varied more lexically than semantically. We found an average and median Jaccard similarity score of 0.0367 and 0, respectively, and an average and median cosine similarity score of 0.1589 and 0, respectively.

We conservatively consider a keyword to be semantically divergent across parties if it has a Jaccard similarity of less than 0.1 and a cosine neighborhood similarity below 0.3, indicating low lexical and conceptual overlap between democrat and republican framings. All but four keywords fall below the 0.1 Jaccard threshold– *education, economy, freedom*, and *veteran*.

Keywords we identified as most divergent are (bolded indicates a Jaccard and cosine similarity score of 0): ***abortion***, *border*, ***capitalism***, ***climate change***, *crime*, *democracy*, ***environment***, *gas*, ***gun***, *healthcare*, ***immigrant***, ***immigration***, ***inflation***, *job*, ***justice***, ***medicare***, *military*, ***police***, ***racism***, ***security***, ***tax***, ***terrorism***, and ***voting right***.

Keywords we identified as most similar are (Jaccard similarity > 0.1 or cosine similarity > 0.4): *education, economy, equality, freedom, military, patriotism, right, truth*, and *veteran*.

This implies that the first list contains polarizing issues and ideas that are used very differently by party, while the second list contains words democrats and republicans frame similarly.

Keyword	Jaccard	Cosine
abortion	0.000	0.000
border	0.053	0.275
capitalism	0.000	0.000
climate_change	0.000	0.000
crime	0.053	0.296
democracy	0.053	0.271
education	0.111	0.399
economy	0.176	0.302
equality	0.053	0.424
environment	0.000	0.000
freedom	0.111	0.378
gas	0.053	0.278
gun	0.000	0.000
healthcare	0.053	0.191
immigrant	0.000	0.000
immigration	0.000	0.000
inflation	0.000	0.000
job	0.053	0.272
justice	0.000	0.000
medicare	0.000	0.000
military	0.053	0.440
patriotism	0.053	0.427
police	0.000	0.000
racism	0.000	0.000
right	0.053	0.405
security	0.000	0.000
social_security	0.053	0.310
tax	0.000	0.000
terrorism	0.000	0.000
truth	0.053	0.405
veteran	0.176	0.431
voting_right	0.000	0.000

Table 2: Jaccard and Cosine Similarities for Political Keywords

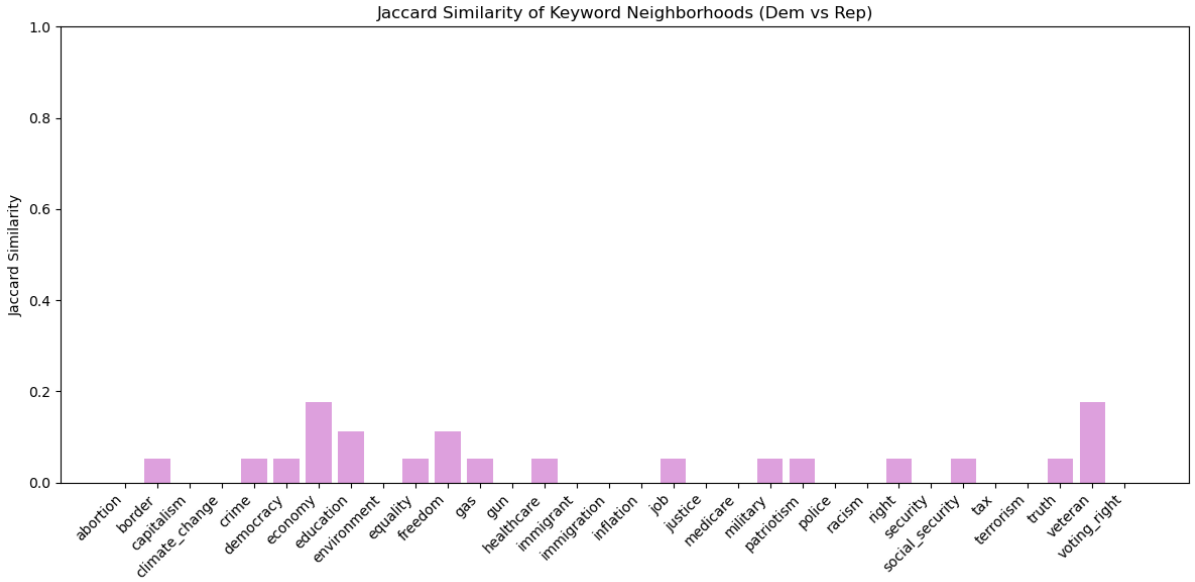

Figure 3: Jaccard Similarity of Keyword Neighborhoods between Democrats and Republicans

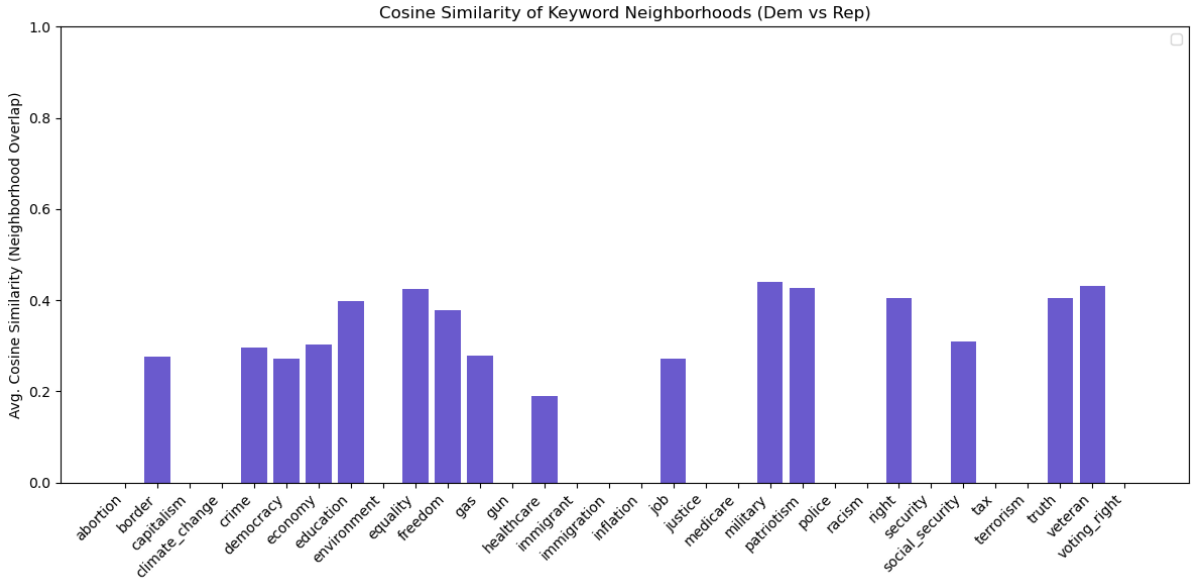

Figure 4: Cosine Similarity of Keyword Neighborhoods between Democrats and Republicans

References

Dallas Card, Serina Chang, Chris Becker, Julia Mendelsohn, Rob Voigt, Leah Boustan, Ran Abramitzky, and Dan Jurafsky. 2022. Computational analysis of 140 years of us political speeches reveals more positive but increasingly polarized framing of immigration. *Proceedings of the National Academy of Sciences*, 119(31):e2120510119.

Jinghui Chen, Takayuki Mizuno, and Shohei Doi. 2024. Analyzing political party positions through multilanguage twitter text embeddings. *Frontiers in Big Data*, Volume 7 - 2024.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Limitations

There are important limitations to take into account. This twitter data is from 2021. Due to the volatility of politics and party messaging, it may not be accurate to the framings of these concepts by each party today. It's also important to note the source of these tweets is just senators, and thus does not capture the depth or range of each political party. There is much framing data from representatives in the house, the president, and in state and local offices that is not represented here. I would encourage future analyses of this kind to expand with larger and more updated datasets. I also would implore further research into the adjustment of framings over time, as in (Card et al., 2022).

There also may be bias, due to extremist senators who may sway the embeddings in ways that are not the standard.

It is also limiting that much of our analysis was qualitative. These observations are more ambiguous, and can be interpreted in many different ways.