You Have the Floor: A Speaker-Aligned Corpus Derived from the Congressional Record

Jennifer L Bochenek, Jake Ryland Williams

Drexel University
jlb599@drexel.edu, jw3477@drexel.edu

Abstract

The *United States Congressional Record* serves as a comprehensive archive of legislative discourse, yet its sheer volume and unstructured format pose significant challenges for researchers interested in analyzing political language, speaker behavior, and ideological framing. This paper presents a new dataset that organizes Congressional speeches by individual speakers. Data was obtained using a mixture of the Congress.gov Application Programming Interface (API) and web-scraping techniques to retrieve the full text of the *Congressional Record*. After extracting roll-call votes and standardizing the transcripts to remove noisy artifacts and normalize formatting, each speaker is separated into individual files and annotated with metadata including name, political affiliation, years active, district or state represented, and professional social media accounts, if known. This enables fine-grained analysis of rhetorical patterns and linguistic strategies across different political groups as well as time periods. By making the dataset publicly available, we aim to support interdisciplinary research utilizing natural language processing (NLP).

Datasets — https://doi.org/10.7910/DVN/5BYHIV

Introduction

The *Congressional Record* provides a comprehensive record of the proceedings and debates of the United States Congress, serving as a crucial resource for understanding legislative behavior, policymaking, and political discourse. Over time, congressional speeches have evolved in response to changing institutional norms, technological advancements, and political dynamics. Analyzing these patterns can reveal critical insights into how legislators communicate, how discourse is shaped by demographic and procedural factors, and how polarization influences political debate.

Existing tools and datasets, such as Judd et al. (2017) and Gentzkow, Shapiro, and Taddy (2018), have laid the groundwork for parsing and analyzing congressional text. These resources have provided access to parsed speeches, phrase counts, and basic metadata; enabling research into historical trends and the linguistic characteristics of congressional discourse (Card et al. 2022; Rodriguez and Spirling 2022).

This work builds on prior efforts by introducing an expanded dataset covering speeches from 1995 to 2024 and integrating comprehensive metadata, including social and biographical information for legislators. It addresses critical challenges in speaker attribution, handling of role-based speakers, and text pre-processing, ensuring high accuracy and usability for downstream analyses. Beyond the descriptive analysis presented in this paper, this work also outlines novel future avenues of research, such as text alignment strategies and text style transfer, to bridge linguistic divides between political parties and depolarize political texts, or for co-reference resolution or entity linking based research into political domains.

The primary contributions of this work are as follows:

- An expanded and up-to-date dataset of congressional speeches enriched with detailed metadata for each legislator.
- Advanced speaker attribution techniques and incorporating role-based speakers.
- Enhanced metadata links linguistic differences that span political parties and individual speakers.

This work aims to contribute a dataset to deepen understanding of congressional discourse and eventually offer practical solutions for addressing polarization in political communication. The dataset can also be used as a named entity rich resource for co-reference resolution and entity linking research.

Dataset Creation

Creating the overall data set involved a number of data acquisition and pre-processing activities, including: data collection and standardization, as well as speaker detection and matching that handles unmatched and role-specific speakers.

Data Collection

The data was collected from the *Congressional Record*, the official record of the proceedings and debates of the United States Congress. It is published daily when Congress is in session. The *Congressional Record* began publication in 1873, but digital versions are available on Congress.gov for 1995 to the present (Congress 2025). The *Congressional Record* consists of four sections:

- **Daily Digest**: A summary of floor and committee activities.
- **House Section**: Proceedings from the House of Representatives.
- **Senate Section**: Proceedings from the Senate.
- **Extension of Remarks**: Tributes, statements, and other supplemental content related to House proceedings.

The metadata for the *Congressional Record* was accessed through the Congress.gov API (Library of Congress 2025). Calls to the `daily-congressional-record` endpoint fetched metadata, which was then used to generate subsequent calls for the URL of the text file for each "article"–a discrete subsection of the Record associated with a specific topic. Once a list of article URLs was compiled for each day, the URLs were traversed, and the corresponding text files were downloaded with a one-second delay between requests. From 1995 to 2024, the dataset covered 5,502 days and consisted of 765,185 articles. The total data retrieval process took approximately two weeks.

After gathering articles for each day, the *Congressional Record* was reconstructed in order using the section and page numbers embedded in each article. Articles could span multiple pages or comprise a subset of a page and represented complete discussions or proceedings. Given the scope of the data already collected, records from earlier years (available only as PDFs or images) were excluded for this project.

Standardization

The next step was standardizing the daily *Congressional Record* files. The standardization process involved removing roll-call votes and fixing character encoding. Roll-call vote results were removed because they were highly structured (i.e., presented in text as long lists of names in alphabetical order grouped by vote type). This could cause issues with eventual model building. Additionally, there were some minor encoding issues that were fixed in the process of standardization to ensure integrity of the text.

Speaker Detection and Matching

One large challenge involved detecting and separating speaker statements and matching them with current or former legislators. Legislator metadata was sourced from the publicly available *congress-legislators dataset* (@unitedstates Project 2025). This dataset includes `legislators-current`, `legislators-historical`, `legislators-social-media`, and `executive` files. These files were merged into a comprehensive database of all terms of office of members of both the legislative and executive branches. Executive members, such as the Vice President, were included due to their speaking roles in Congress. The entirety of the @unitedstates Project is dedicated to the public domain and copyright and related rights are waived, so can be used in part or entirety for this dataset.

Speakers were detected using regex based on the standard formula used in the *Congressional Record*: "Mr./Mrs./Ms./Miss. LASTNAME." sometimes followed by "of [State]." (e.g., Mr. ABERNETHY or Mr. ABERNETHY of North Carolina). Once a speaker was identified, all subsequent text was attributed to them until a new speaker was detected or the end of the article was reached.

In order to match speakers to legislators:

1. **Date Filtering**: Only legislators serving at the time of record entry were considered.
2. **Chamber Filtering**: Matches were further refined by chamber (House or Senate) based on the section/page number (i.e., "S" for Senate, "H" for House, and "E" for Extension of Remarks, which exclusively pertains to the House).
3. **Gender Filtering**: Gender was identified using speaker title in the text (i.e., Mr. for congressmen and Mrs./Ms./Miss. for congresswomen).
4. **Matching and Validation**: Speaker names were matched to the filtered list of legislators. Each speech could only have one match; texts with more than one match or no matches were treated separately.

Unmatched Speakers For those speakers who could not be matched, their speech data was included in a separate "unmatched speakers" file. Common reasons for unmatched speakers included:

- **Non-Legislators**: Invited speakers or external participants.
- **Ambiguous Matches**: Cases where more than one legislator matched (e.g., "Mr. NELSON" could refer to multiple individuals). Ambiguous matches were flagged and logged with all potential bioguide IDs for later review. The matching bioguide IDs were appended to the first line of the speech text between brackets if disambiguation is later desired.
- **Typographical Errors**: Errors in recording a speaker's name (e.g., missing or extra letters). Although fuzzy matching could resolve some cases, it risked false positives. To avoid misattribution, these cases were also flagged as unmatched.

An example of ambiguous matches is Mr. DAVIS from 2003. The name, date, and chamber matched to five distinct individuals in the House of Representatives: D000096, Danny Davis of Illinois (D); D000114, Jim Davis of Florida (R); D000136, Tom Davis of Virginia (R); D000602, Artur Davis of Alabama (D); or D000599, Lincoln Davis of Tennessee (D). There are 1,085 unmatched speakers with 4,210 speeches total.

Role Speakers In addition to unmatched speakers, there are speeches that are not aligned with a specific person but instead with a role they assume in managing political discourse and the flow of legislation. These role-based attributions are an integral part of the legislative process and often represent procedural statements, rulings, or guidance during debates.

The special cases include the following roles:

- **The CHAIR/The Acting CHAIR**: A presiding officer over a session in the House, Senate, or a specific committee. Responsibilities include managing debates, enforcing rules of order and ensuring proper legislative procedures are followed. The title can also be used generically to refer to various leadership roles in the legislative process. (Heitshusen 2019; U.S. House of Representatives 2025)
- **The CHAIRMAN/The Acting CHAIRMAN/Mr. CHAIRMAN**: The chairman leads a congressional committee. Responsibilities include setting committee's agenda, managing hearings and investigations, and guiding legislation through the committee process. (U.S. House of Representatives 2025)
- **The PRESIDENT/The ACTING PRESIDENT pro tempore**: The Vice-President acting in their constitutional role as President of the Senate or their delegate. Responsibilities include presiding over Senate sessions, breaking ties in votes, and maintaining procedural order. (Heitshusen 2019)
- **The PRESIDING OFFICER**: A general term for the role of the person managing the proceedings in the House or Senate at any given time. Responsibilities include enforcing rules of order, managing debates and voting processes, and ensuring that legislative sessions progress efficiently. This role can be filled by various officials, including the Speaker/Speaker pro tempore or their delegates. (Gamm and Smith 2000)
- **The SPEAKER/The SPEAKER pro tempore/Mr. SPEAKER**: The presiding officer of the House of Representatives. Responsibilities include presiding over house sessions, setting legislative agenda, representing the House in ceremonial roles, and assigning bills to committees. (Heitshusen 2019)

These roles are used in place of names in the *Congressional Record* and thus do not correspond to a specific individual but rather to the official capacity held by the speaker at the time. To preserve the integrity of the data, these entries were categorized under their respective roles rather than attempting to match them to individual legislators. There are 601,306 speeches across the aforementioned roles.

By tagging the above entries with their role, the dataset retains its structure while providing insight into procedural discourse. These role speakers were handled separately during pre-processing and included in the dataset as distinct entities, ensuring that their contributions to the legislative process are fully captured.

Dataset Formatting

To organize the processed data effectively, JSON files were created for each legislator, with the filename corresponding to the legislator's Bioguide ID. This format was chosen for its flexibility and compatibility with various data analysis tools. Each JSON file contains comprehensive metadata about the legislator and their associated speeches.

JSON Structure

The structure of a typical JSON file is as follows:


```
{
  'bioguide': 'D000563',
  'display_name': 'Mr. DURBIN',
  'display_name_with_state': 'Mr. DURBIN of Illinois',
  'name': {
    'first': 'Richard',
    'middle': 'J.',
    'last': 'Durbin',
    'official_full': 'Richard J. Durbin'
  },
  'id': {
    'bioguide': 'D000563',
    'thomas': '00326',
    'lis': 'S253',
    'govtrack': 300038,
    'opensecrets': 'N00004981',
    'votesmart': 26847,
    'fec': [
      'S6IL00151',
      'H2IL20026'
      ],
    'cspan': 6741,
    'wikipedia': 'Dick Durbin',
    'house_history': 12527,
    'ballotpedia': 'Dick Durbin',
    'maplight': 563,
    'icpsr': 15021,
    'wikidata': 'Q434804',
    'google_entity_id': 'kg:/m/01xcd1'
  },
  'bio': {
    'birthday': '1944-11-21',
    'gender': 'M'
  },
  'terms': [
    {
      'type': 'rep',
      'start': '1983-01-03',
      'end': '1985-01-03',
      'state': 'IL',
      'party': 'Democrat'
    }, ...
  ],
  'social': {
    'twitter': 'SenatorDurbin',
    'facebook': 'SenatorDurbin',
    'youtube': 'SenatorDurbin',
    'youtube_id': 'UCkbixlNCxcKAffEhe3X5-lw',
    'twitter_id': '247334603'
  },
  'speeches': [
    {
      'date': '1995-06-20',
      'party': 'Democrat',
      'chamber': 'house',
      'text': 'Mr. Speaker, today I am introducing
      legislation along with Representative Dave Camp
      to encourage organ donation through a highly
      cost-effective campaign of public education...'
      'speech_id': 'D0005631995-06-2024',
      'previous_speech_id': 'D0005631995-06-2024',
    },
    {
      'date': '1995-06-22',
      'party': 'Democrat',
```

```
      'chamber': 'house',
      'text': 'Mr. Speaker, I want to thank the
      gentlewoman from Texas for taking this special
      order...'
      'speech_id': 'D0005631995-06-22506',
      'previous_speech_id': 'J0000321995-06-22505',
    },
    ...
  ]
}
```

The explanation for the various components of the JSON structure is as follows:

- **Bioguide ID**: A unique identifier for each legislator.
- **Name Information**: Full name; first, middle, and last names; and the names used for merging - display name and display name with state.
- **Identifiers**: Cross-references to external systems, including GovTrack, OpenSecrets, and Wikidata.
- **Biography**: Details such as date of birth and gender.
- **Terms**: Service periods, including type (e.g., representative or senator), start/end dates, state, and party affiliation.
- **Social Media**: Handles and IDs for platforms like Twitter, Facebook, and YouTube, if available.
- **Speeches**: A list of speeches attributed to the legislator, including date, party, chamber, text, speech ID, and previous speech ID (the same as speech ID if there is no preceding speech). Each speech ID was built using the speaker's bioguide ID, if available, or name if not, then the date of the speech and finally the speech index.

Handling Unmatched and Role-Based Speakers

Datasets for unmatched speakers and role-based entries (e.g., "The SPEAKER" or "The PRESIDING OFFICER") follow a similar structure to the `speeches` section in the above JSON structure. However, these entries lack personal metadata (e.g., biographical or identifier details) and are instead categorized under their respective roles or names. This ensures comprehensive coverage of all discourse, even when a specific legislator cannot be attributed.

Review of Data Collection

The dataset includes legislators from five distinct political affiliations: Democrat, Republican, Independent, Libertarian, and Popular Democrat. Below are the basic counts of the party representation:

Major and Minor Parties

The two major parties, Democrat and Republican, dominate the dataset with:

- **Democrats**: 730 legislators
- **Republicans**: 861 legislators

In addition to these, the dataset includes legislators from smaller affiliations:

- **Popular Democrat**: 1 legislator
- **Libertarian**: 1 legislator
- **Independent**: 11 legislators

Corrections and Manual Edits

During data preparation, a manual correction was applied to the entry for **A000359** (Aníbal Acevedo Vilá). His party affiliation for the 107th Congress was incorrectly listed as `AL` in the raw data. This was corrected to `Popular Democrat`, as he represented Puerto Rico and is the only legislator in the dataset with this affiliation.

Independent and Minor Party Legislators

The dataset includes the following independent and minor party legislators:

- **Popular Democrat**: Aníbal Acevedo Vilá (`A000359`) of Puerto Rico.
- **Libertarian**: Justin Amash (`A000367`) of Michigan, previously a Republican.
- **Independents** (11 legislators):
  - Dean Barkley (`B001237`) of Minnesota
  - Victor Frazer (`F000351`) of the U.S. Virgin Islands
  - Virgil Goode (`G000280`) of Virginia
  - Jim Jeffords (`J000072`) of Vermont
  - Angus King (`K000383`) of Maine
  - Joe Lieberman (`L000304`) of Connecticut
  - Joe Manchin (`M001183`) of West Virginia
  - Paul Mitchell (`M001201`) of Michigan
  - Bernie Sanders (`S000033`) of Vermont
  - Gregorio Sablan (`S001177`) of Northern Mariana Islands
  - Kyrsten Sinema (`S001191`) of Arizona

For simplicity in initial reporting, the data were reduced to two parties (Democrat and Republican) for the remainder of this paper.

Total Legislators and Cross-Affiliations

The dataset includes a total of 1,589 unique entries when examining legislators by party affiliation. This total exceeds the number of distinct legislators because some individuals were affiliated with more than one party during their careers. For instance:

- **Joe Manchin** (`M001183`), initially a Democrat, later identified as an Independent.
- **Justin Amash** (`A000367`), who switched from Republican to Libertarian.

This overlap demonstrates the complexity of party affiliation over time and highlights the need for detailed metadata tracking to account for such changes.

The description of the dataset by party, chamber, gender, and year are presented in Table 1. Additionally, figure 1 shows the number of speakers, speeches, and words by state after being normalized by the number of speakers.

		Total Speakers		Total Speeches			Total Words			
		N	%	N	%	Per Speaker	N	%	Per Speaker	Per Speech
Party	Democrat	730	45.88%	715,521	50.76%	980.02	299,697,762	52.00%	410,897.48	419.27
	Republican	861	54.12%	693,349	49.24%	806.08	275,875,530	48.00%	321,553.46	398.91
Chamber	House	1380	83.48%	922,208	65.43%	668.27	319,936,796	55.46%	231,838.26	346.92
	Senate	273	16.52%	487,244	34.57%	1,784.77	256,875,897	44.54%	940,937.35	527.20
Gender	Female	293	18.47%	198,585	14.09%	677.76	81,877,944	14.19%	279,446.91	412.31
	Male	1293	81.53%	1,210,867	85.91%	936.47	494,934,749	85.81%	382,780.16	408.74
Year	1995	534	3.39%	98,256	6.97%	184.00	35,663,011	6.18%	66,784.66	362.96
	1996	533	3.39%	60,257	4.28%	113.05	24,022,415	4.16%	45,070.20	398.67
	1997	538	3.42%	60,865	4.32%	113.13	26,459,084	4.59%	49,180.45	434.72
	1998	535	3.40%	57,964	4.11%	108.34	23,381,476	4.05%	43,703.69	403.38
	1999	537	3.41%	65,335	4.64%	121.67	28,956,005	5.02%	53,921.80	443.19
	2000	535	3.40%	58,826	4.17%	109.96	25,006,761	4.34%	46,741.61	425.10
	2001	539	3.42%	54,424	3.86%	100.97	24,114,585	4.18%	44,739.49	443.09
	2002	531	3.37%	47,572	3.38%	89.59	20,084,518	3.48%	37,823.95	422.19
	2003	527	3.35%	62,241	4.42%	118.10	27,585,966	4.78%	52,345.29	443.21
	2004	526	3.34%	49,669	3.52%	94.43	21,542,883	3.73%	40,956.05	433.73
	2005	525	3.34%	57,276	4.06%	109.10	26,014,476	4.51%	49,551.38	454.20
	2006	520	3.30%	46,520	3.30%	89.46	20,374,563	3.53%	39,181.85	437.97
	2007	523	3.32%	69,533	4.93%	132.95	27,767,055	4.81%	53,091.88	399.34
	2008	522	3.32%	47,171	3.35%	90.37	20,489,655	3.55%	39,252.21	434.37
	2009	528	3.35%	65,308	4.63%	123.69	25,592,333	4.44%	48,470.33	391.87
	2010	525	3.34%	45,220	3.21%	86.13	17,867,383	3.10%	34,033.11	395.12
	2011	528	3.35%	48,780	3.46%	92.39	17,964,980	3.11%	34,024.58	368.29
	2012	527	3.35%	37,531	2.66%	71.22	13,876,913	2.41%	26,331.90	369.75
	2013	541	3.44%	40,170	2.85%	74.25	15,663,103	2.72%	28,952.13	389.92
	2014	520	3.30%	35,033	2.49%	67.37	13,291,228	2.30%	25,560.05	379.39
	2015	514	3.27%	39,935	2.83%	77.69	14,895,534	2.58%	28,979.64	372.99
	2016	511	3.25%	31,794	2.26%	62.22	11,906,831	2.06%	23,301.04	374.50
	2017	509	3.23%	37,095	2.63%	72.88	16,402,498	2.84%	32,224.95	442.18
	2018	514	3.27%	31,214	2.21%	60.73	13,434,663	2.33%	26,137.48	430.41
	2019	514	3.27%	34,507	2.45%	67.13	12,573,571	2.18%	24,462.20	364.38
	2020	511	3.25%	23,614	1.68%	46.21	11,145,569	1.93%	21,811.29	471.99
	2021	516	3.28%	25,950	1.84%	50.29	11,903,681	2.06%	23,069.15	458.72
	2022	518	3.29%	24,650	1.75%	47.59	9,766,201	1.69%	18,853.67	396.19
	2023	518	3.29%	26,971	1.91%	52.07	9,034,937	1.57%	17,441.96	334.99
	2024	521	3.31%	25,771	1.83%	49.46	10,030,815	1.74%	19,253.00	389.23

Table 1: Description of the dataset, by party, chamber, gender, and year as well as by number of speakers, total number of speeches, and total words.

Discussion

The dataset reveals several notable trends and patterns concerning the number of words spoken, the length of speeches, and the distribution of speeches by demographic and geographic factors. These trends highlight evolving patterns in congressional discourse and provide insights into procedural, demographic, and institutional dynamics.

Decline in Words per Speaker and Number of Speeches Over the years, the average number of words spoken per speaker has declined significantly (Table 1). This decline coincides with a reduction in the total number of speeches given, despite the average number of words per speech remaining relatively constant (Table 1). These patterns suggest that while individual speeches are not getting shorter, fewer legislators are speaking overall.

This trend may reflect structural changes in Congress, such as stricter time limits on debate, the increased use of unanimous consent agreements, and shifts in legislative priorities (Evans and Oleszek 2000). It may also result from heightened political polarization, which has been shown to reduce bipartisan deliberation and increase the use of prewritten or scripted remarks (Henry 2013).

Chamber-Specific Trends in Speech Length The data shows that senators consistently speak more words per speech (527.20 on average) than representatives in the House (346.92) (Table 1 and 4). Senators also speak more words per speaker (Table 1 and Figure 5), with an average of 940,937.35 compared to 231,838.26 in the House. The Senate has not been spared the decline in words per speaker over time, and seems to be falling faster than the House of

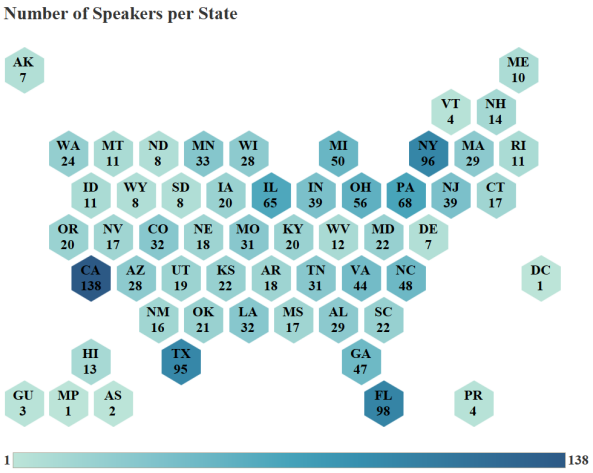

Figure 1: Number of speakers by state.

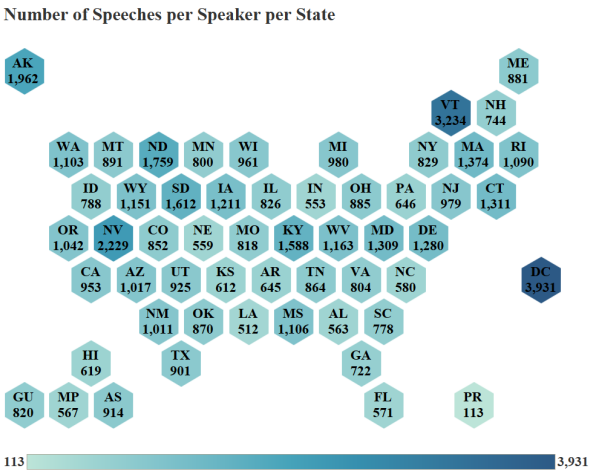

Figure 2: Number of speeches per speaker by state.

Representatives in that regard (Figure 5. These patterns align with the Senate's smaller membership and its role as a deliberative body, where extended debate is encouraged (Smith 2014; Evans and Oleszek 2000). In contrast, the House operates under stricter rules to manage time and ensure broader participation among its larger membership.

Geographic Trends and Distribution The geographic distribution of speakers reveals a predictable correlation between state population size and the number of speakers, particularly in the House of Representatives (Figure 1). Larger states such as California and Texas contribute more speakers, reflecting their larger House delegations. However, the distribution across states remains relatively equitable when considering the Senate, where each state has equal representation.

However, there are interesting results when examining the data by number of speeches per speaker (Figure 2) and by number of words in speeches per speaker by state (Figure 3. Using these metrics, the lead for the higher represented states evens out (given the relative color uniformity in the graphs), while notable outliers include Vermont and Washington, D.C. Vermont has the highest number of speeches per speaker (3,234) and words per speaker (2,238,541) (Figures 2 and 3). This trend is likely due to its small delegation and the prominence of individual legislators, such as Bernie Sanders, who are known for their frequent and lengthy speeches. Washington, D.C., meanwhile, shows a high number of words per speaker (1,669,566) despite limited voting representation, reflecting its unique political role. Other notable states include Alaska, Nevada, and North Dakota.

Party and Gender Representation The dataset reflects a relatively even split between Democratic and Republican contributions to the total number of speeches and words spoken (Table 1, Figures 6 & 7). However, Democrats tend to give slightly longer speeches on average (419.27 words per speech compared to 398.91 for Republicans). This could reflect differences in rhetorical styles or procedural roles, as suggested by Aldrich and Rohde (2000) and Lee (2016), who note that party leaders often set the tone for debate.

However, the representation of gender shows a marked disparity. Female legislators account for only 18.47% of speakers and 14.09% of speeches, reflecting ongoing underrepresentation (Table 1). Despite this, female legislators tend to deliver slightly longer speeches (412.31 words per speech) than their male counterparts (408.74) (Figures 8 & 9 show trends over time). This aligns with findings by Pearson and Dancey, who argue that women often maximize their speaking opportunities to establish authority and influence in male-dominated institutions (2011). Furthermore, Miller and Sutherland (2023) demonstrated that congresswomen are more likely than congressmen to be interrupted during committee hearings—up to twice as often when discussing issues that impact women's rights. These patterns suggest that female legislators face unique challenges in legislative discourse, which may encourage them to use their speaking opportunities strategically to assert influence and advocate for underrepresented issues.

Instability in Speech-making Patterns over Time Despite the decline in the number of speeches (Table 1, column Total Speeches and Total Speeches per Speaker), the average number of words per speech has remained relatively stable over time (Table 1, column Total Words per Speech). This stability suggests that, while fewer legislators are speaking, those who do speak maintain similar verbosity. This may reflect evolving norms in legislative speech-making or procedural changes.

Implications for Future Research

These findings highlight the evolving nature of congressional discourse and open avenues for further research. Potential areas of exploration include:

- The impact of procedural changes, such as unanimous consent agreements and stricter time limits, on speech patterns.

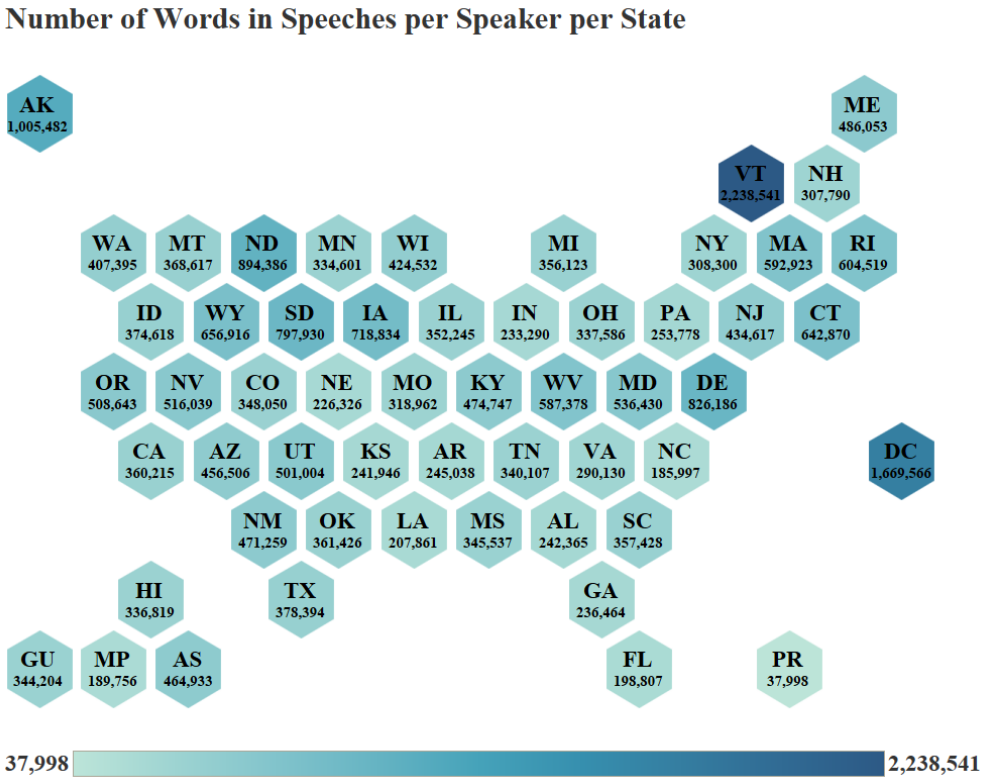

Figure 3: Number of words in speeches per speaker by state.

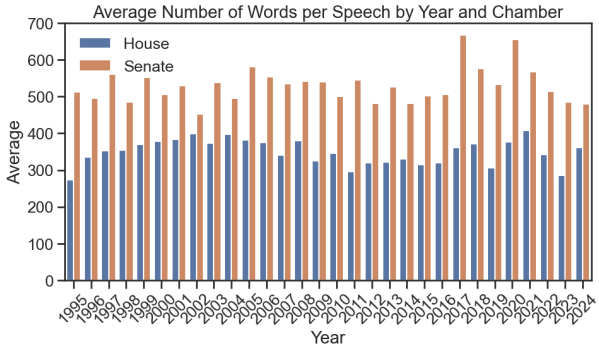

Figure 4: The average number of words spoken by year in each chamber of Congress.

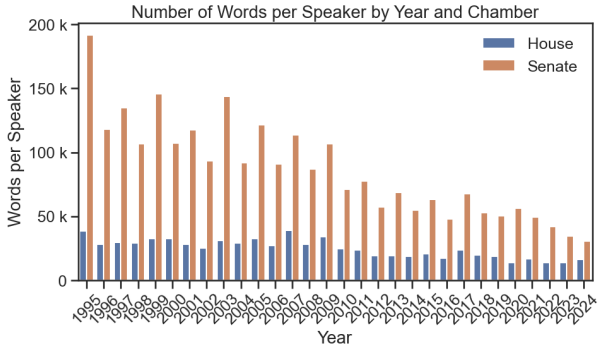

Figure 5: Number of words spoken per speaker by year in each chamber of Congress.

- The role of political polarization in reducing bipartisan debate and increasing scripted or symbolic speeches.
- The influence of demographic factors, such as gender and geographic representation, on legislative discourse.
- The unique speech patterns of outliers like Vermont and Washington, D.C., and their broader implications for legislative norms.

Future studies could also examine the qualitative content of speeches to understand how these trends influence policymaking and public perception of Congress.

The specific purview of the author lies in language differences between major political parties, particularly in the areas of text alignment strategies and text style transfer. This eventual research aims to depolarize political texts by identifying linguistic patterns that contribute to polarization and proposing methods to bridge stylistic divides. The dataset presented here will serve as a foundation for these efforts, enabling the development of algorithms and tools to analyze and modify speech patterns in a way that fosters bipartisan communication and reduces ideological divides.

Additionally, the dataset will also be enhanced through

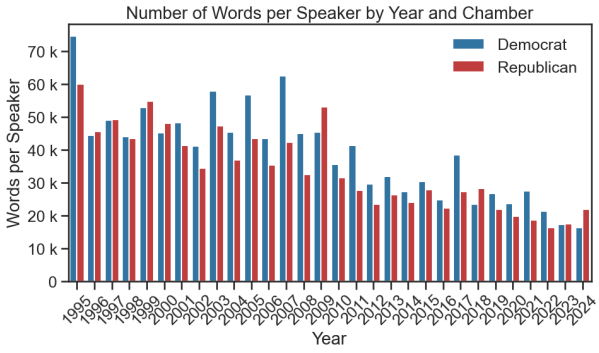

Figure 6: Number of words spoken per speaker by year between the two major political parties.

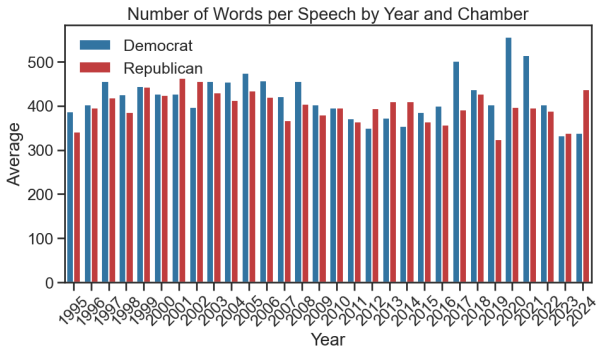

Figure 7: The average number of words spoken by year between the two major political parties.

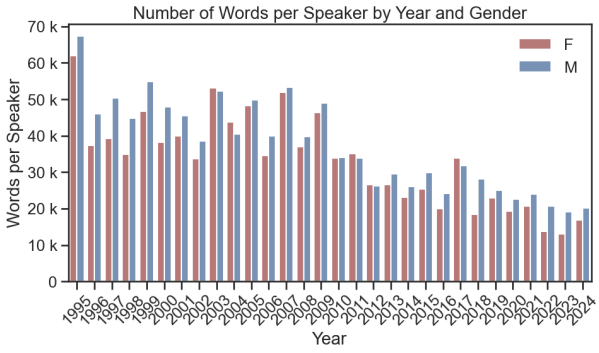

Figure 8: Number of words per speaker by year split between congressmen and congresswomen.

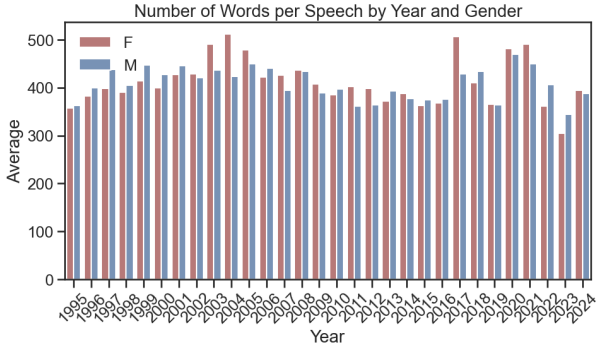

Figure 9: The average number of words spoken by year split between congressmen and congresswomen.

the application of co-reference resolution and entity linking techniques, which will provide greater contextual coherence and semantic richness (Samarawickrama et al. 2020). Co-reference resolution will help identify when different expressions in the text (e.g., "the Senator," "she," "Dr. Smith") refer to the same entity, improving the accuracy of speaker attribution and discourse analysis (Zhukova et al. 2021). Meanwhile, entity linking will connect mentioned individuals, organizations, or locations in the speeches to structured knowledge bases such as Wikidata or information on bills and proposals (Erjavec et al. 2023). Together, these enhancements will enable more sophisticated analyses of rhetorical patterns, topic continuity, and the roles of referenced actors throughout congressional discourse.

Beyond the previously stated applications, this dataset offers substantial utility across multiple domains. Political scientists could use speaker-aligned data to study changes in legislative rhetoric over time or to track discourse around specific policy areas. Sociologists may find value in exploring speech patterns by gender, region, or political affiliation. For computational linguists, the dataset provides a large, labeled corpus suitable for tasks such as stance detection, topic modeling, or rhetorical strategy classification. The dataset may also support journalistic or civic research; for instance, enabling the tracking of how certain issues emerge and evolve across sessions or how individual legislators shift in tone or focus. By linking speech content to speaker identity and contextual metadata, this dataset offers broad utility for analyzing institutional discourse and political communication.

Comparison to Previous Works

As mentioned, there have been some previous attempts to process the *Congressional Record* into a dataset for research purposes. After describing the dataset collected for this paper, it is important to discuss how it differs from these extent datasets.

The output of Judd et al. (2017) is a python-based command line interface (CLI) which fetches and process the *Congressional Record* on demand. This is not a ready-made dataset of the *Congressional Record*, and requires sufficient computational skills and computational power to get the CLI installed and working. Additionally, like the other example Gentzkow, Shapiro, and Taddy (2018), the data are speech focused, rather than speaker focused. The output is a json file per article within a folder for each day, nested in a folder per year, with metadata appended from a source similar to the @unitedstates project (@unitedstates Project 2025). This CLI is limited to data from 1994 to current day.

The Gentzkow, Shapiro, and Taddy (2018) dataset is similarly speech focused, comprised of text files for each seating Congress from the 97th Congress (starting January 3rd,

1981) to the 114th (ending January 3rd, 2017). Each line in the text file is a new speech, cross-referenced to another text file with minimal speaker metadata. The data also excludes the Extension of Remarks section of the *Congressional Record* and separates out special roles with unmatched speakers together (including failing to correctly parse several speakers).

While the dataset for this paper is presented in a speaker-aligned manner, it could be re-aligned to a dialog-based dataset or speech order-based dataset. As each speech is labeled with a unique ID and the ID of the previous speech, it would be relatively easy to reconstruct the dataset to examine conversational or turn-taking legislative discussions. Date and speech order data are also embedded in the speech ID, allowing for a data structure similar to Gentzkow, Shapiro, and Taddy (2018) if needed.

Limitations

While this dataset provides relatively comprehensive data for analyzing recent congressional discourse, there are several limitations and ethical considerations to acknowledge:

- **Date limitations**: This dataset only covers speeches from 1995 to 2024, leaving out earlier, bound, copies of the *Congressional Record* that could potentially provide additional value into insights and increase the size of data for each party. A reason to maintain the more limited scope is that Party positions on issues do drift over time, so early text from members of the Democrat or Republican parties might not be aligned with current party lines.
- **Speaker attribution errors**: Although every effort was expended to attribute speakers to the correct legislator, some ambiguities do remain. Thus, there are members of the unmatched speakers file that could be matched manually at a later date.
- **Dependence on accuracy of metadata**: The quality of the matching requires accurate external metadata, and a manual correction to the metadata was noted earlier in the paper. There could be other issues with the metadata that are not caught and may need to be updated at a later date.

Conclusions

This paper presents a comprehensive speaker-aligned dataset derived from the *United States Congressional Record*, covering speeches from 1995 to 2024. By pairing speeches with detailed metadata–including speaker demographics, political affiliations, and role-based attributions–this dataset offers researchers a robust resource for exploring legislative discourse. Using detailed text processing methods, such as speaker attribution and role identification, this work addresses key challenges in parsing large-scale legislative text.

The analysis of the dataset reveals important trends in congressional discourse, including the decline in words spoken per speaker, disparities in gender representation, and the unique verbosity of speakers associated with specific states. These findings underscore the evolving dynamics of legislative communication, shaped by institutional norms, demographic factors, and political polarization. The dataset enables nuanced analyses of these trends, providing a foundation for addressing critical questions in political science, computational linguistics, and social research.

The data's speaker-focused structure opens avenues for studying individual rhetorical strategies, partisan discourse, and demographic influences on legislative behavior. Future research is planned to leverage this dataset to investigate text alignment strategies, text style transfer, and other techniques aimed at reducing political polarization. Additionally, the dataset is going to be used for other avenues of research such as named entity detection and co-reference resolution and entity linking in political domains. Beyond these stated applications, this dataset could contribute to ongoing efforts to democratize access to legislative data and foster interdisciplinary collaboration.

However, this work is not without limitations. The exclusion of data before 1995, the presence of unmatched speakers, and potential errors in metadata highlight the need for cautious interpretation. Moreover, while the dataset aims to promote constructive research, it is imperative to guard against its misuse, particularly in contexts that could perpetuate stereotyping, misinformation, or ideological division.

By addressing these challenges and building upon the findings presented here, researchers can deepen their understanding of congressional discourse and its broader societal implications. In doing so, this work serves as a stepping stone toward depolarizing or discovering co-references in political communication and fostering a more informed and engaged public.

References

Aldrich, J. H.; and Rohde, D. W. 2000. The consequences of party organization in the House: The role of the majority and minority parties in conditional party government. In *Polarized politics: Congress and the president in a partisan era*, volume 31, 33–34. CQ Press. Accessed: 2025-01-15.

Card, D.; Chang, S.; Becker, C.; Mendelsohn, J.; Voigt, R.; Boustan, L.; Abramitzky, R.; and Jurafsky, D. 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. Accessed: 2025-01-15.

Congress, U. 2025. Congressional Record. https://www.congress.gov/congressional-record. Accessed: 2024-12-15.

Erjavec, T.; Ogrodniczuk, M.; Osenova, P.; Ljubešić, N.; Simov, K.; Pančur, A.; Rudolf, M.; Kopp, M.; Barkarson, S.; Steingrímsson, S.; et al. 2023. The ParlaMint corpora of parliamentary proceedings. *Language resources and evaluation*, 57(1): 415–448. Accessed: 2025-04-15.

Evans, C. L.; and Oleszek, W. J. 2000. The procedural context of Senate deliberation. In Loomis, B. A., ed., *Esteemed colleagues: Civility and deliberation in the US Senate*, 78–104. Washington DC: Brookings. Accessed: 2025-01-15.

FORCE11. 2020. The FAIR Data principles. https://force11.org/info/the-fair-data-principles/. Accessed: 2025-01-15.

Gamm, G.; and Smith, S. S. 2000. Last among Equals: The Senate's Presiding Officer. In Loomis, B. A., ed., *Esteemed Colleagues: Civility and Deliberation in the U.S. Senate*, 105–134. Washington DC: Brookings. Accessed: 2025-01-15.

Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92. Accessed: 2025-04-15.

Gentzkow, M.; Shapiro, J. M.; and Taddy, M. 2018. Congressional record for the 43rd-114th congresses: Parsed speeches and phrase counts. https://data.stanford.edu/congresstext. Accessed: 2025-01-15.

Heitshusen, V. 2019. Party Leaders in the United States Congress, 1789-2019. Technical Report RL30567, Congressional Research Service. Accessed: 2025-01-15.

Henry, J. 2013. Political polarization: an exploration of its effects on congressional action and public opinion. http://doi.org/10.18297/honors/16. Accessed: 2025-01-15.

Judd, N.; Drinkard, D.; Carbaugh, J.; and Young, L. 2017. Congressional-record: A parser for the Congressional Record. https://github.com/unitedstates/congressional-record/. Accessed: 2025-01-15.

Lee, F. E. 2016. *Insecure majorities: Congress and the perpetual campaign*. University of Chicago Press. Accessed: 2025-01-15.

Library of Congress. 2025. Congress.gov API. https://api.congress.gov. Accessed: 2024-12-15.

Miller, M. G.; and Sutherland, J. L. 2023. The effect of gender on interruptions at congressional hearings. *American Political Science Review*, 117(1): 103–121. Accessed: 2025-01-15.

Pearson, K.; and Dancey, L. 2011. Elevating women's voices in Congress: Speech participation in the House of Representatives. *Political Research Quarterly*, 64(4): 910–923. Accessed: 2025-01-15.

Rodriguez, P. L.; and Spirling, A. 2022. Word embeddings: What works, what doesn't, and how to tell the difference for applied research. *The Journal of Politics*, 84(1): 101–115. Accessed: 2025-01-15.

Samarawickrama, C.; de Almeida, M.; de Silva, N.; Ratnayaka, G.; and Perera, A. S. 2020. Party identification of legal documents using co-reference resolution and named entity recognition. In *2020 IEEE 15th international conference on industrial and information systems (ICIIS)*, 494–499. IEEE. Accessed: 2025-04-15.

Smith, S. S. 2014. *The Senate Syndrome: The Evolution of Procedural Warfare in the Modern US Senate*, volume 12. University of Oklahoma Press. Accessed: 2025-01-15.

@unitedstates Project. 2025. Congress-legislators Dataset. https://github.com/unitedstates/congress-legislators. Accessed: 2024-12-15.

U.S. House of Representatives. 2025. Committee Rules. https://cha.house.gov/committee-rules. Accessed: 2025-01-15.

Zhukova, A.; Hamborg, F.; Donnay, K.; and Gipp, B. 2021. XCoref: Cross-document Coreference Resolution in the Wild. Accessed: 2025-04-15, arXiv:2109.05252.

Paper Checklist

1. For most authors...
   (a) Would answering this research question advance science without violating social contracts, such as violating privacy norms, perpetuating unfair profiling, exacerbating the socio-economic divide, or implying disrespect to societies or cultures? Yes, the collection and sharing of this dataset would not violate any social contracts, privacy norms, or imply disrespect. The data is publicly published for all citizens of the world to access and interpret as desired. This dataset, by making publicly available in a processed form, can advance science by removing barriers of time and computing power to parse the data from scratch.
   (b) Do your main claims in the abstract and introduction accurately reflect the paper's contributions and scope? Yes, the abstract accurately represents the dataset presented in the paper.
   (c) Do you clarify how the proposed methodological approach is appropriate for the claims made? The author(s) do not make any claims outside of a description of the dataset, including how it was collected, processed, and stored.
   (d) Do you clarify what are possible artifacts in the data used, given population-specific distributions? Yes, the Author(s) mention that beyond changes to encoding errors, the text is unaltered. Thus any mistakes, typographical errors, or misattributions are of the *Congressional Record*'s own. This includes mistakes to speaker names that may result in speakers being unmatched.
   (e) Did you describe the limitations of your work? Yes, in the *Limitations and Ethics* subsection.
   (f) Did you discuss any potential negative societal impacts of your work? Yes, in the *Limitations* subsection.
   (g) Did you discuss any potential misuse of your work? Yes, in the *Ethical Statement* subsection.
   (h) Did you describe steps taken to prevent or mitigate potential negative outcomes of the research, such as data and model documentation, data anonymization, responsible release, access control, and the reproducibility of findings? Yes, the data documentation was discussed and how steps were taken to minimize misattribution of the data.
   (i) Have you read the ethics review guidelines and ensured that your paper conforms to them? Yes
2. Additionally, if your study involves hypotheses testing...
   (a) Did you clearly state the assumptions underlying all theoretical results? NA
   (b) Have you provided justifications for all theoretical results? NA

(c) Did you discuss competing hypotheses or theories that might challenge or complement your theoretical results? NA

(d) Have you considered alternative mechanisms or explanations that might account for the same outcomes observed in your study? NA

(e) Did you address potential biases or limitations in your theoretical framework? NA

(f) Have you related your theoretical results to the existing literature in social science? NA

(g) Did you discuss the implications of your theoretical results for policy, practice, or further research in the social science domain? NA

3. Additionally, if you are including theoretical proofs...

(a) Did you state the full set of assumptions of all theoretical results? NA

(b) Did you include complete proofs of all theoretical results? NA

4. Additionally, if you ran machine learning experiments...

(a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? NA, no machine learning was run yet, but all data will be available

(b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? NA

(c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? NA

(d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? NA

(e) Do you justify how the proposed evaluation is sufficient and appropriate to the claims made? NA

(f) Do you discuss what is "the cost" of misclassification and fault (in)tolerance? NA

5. Additionally, if you are using existing assets (e.g., code, data, models) or curating/releasing new assets, **without compromising anonymity**...

(a) If your work uses existing assets, did you cite the creators? Yes, the Author(s) are grateful to the previous work of @unitedstates project for their compilations of legislator metadata that was used to enhance the dataset

(b) Did you mention the license of the assets? Yes, the license is CC0 1.0 Universal public domain for the material from the @unitedstates project, while the *Congressional Record* specifies that —excepting copyrighted articles– there are no restrictions on the republication of the material therein.

(c) Did you include any new assets in the supplemental material or as a URL? Yes, the dataset will be shared as a supplemental material/link to the data

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? NA, all data are a matter of public record.

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? While not explicitly discussed in the paper, the data does contain personally identifiable data on members of Congress, and names and speech data of people who have spoken on the Congress floor. Those who speak in that setting know that they are being recorded and that the matter will be of the public record in perpetuity. The identifying information of members of Congress (e.g., name, birth-date, office address, social media IDs) in the data are part of public record.

(f) If you are curating or releasing new datasets, did you discuss how you intend to make your datasets FAIR (see FORCE11 (2020))? Yes, the FAIR guidelines were discussed broadly in the *Dataset Formatting* section, but to explicate: the dataset uses a standard metadata from (@unitedstates Project 2025) and explained in the *JSON Structure* subsection. The data are stored in a JSON structure that is accessible to many different computational languages, and all source data in the final dataset are public domain. Subsequently, the dataset is available to the public with no restrictions. Thus it meets the FAIR guidelines.

(g) If you are curating or releasing new datasets, did you create a Datasheet for the Dataset (see Gebru et al. (2021))? No, while a separate datasheet for the dataset was not created, most of the sections in Gebru et al. (2021) are also found in this paper.

6. Additionally, if you used crowdsourcing or conducted research with human subjects, **without compromising anonymity**...

(a) Did you include the full text of instructions given to participants and screenshots? NA

(b) Did you describe any potential participant risks, with mentions of Institutional Review Board (IRB) approvals? NA

(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? NA

(d) Did you discuss how data is stored, shared, and deidentified? NA

Ethical Statement

There are potential ethical considerations regarding how this dataset may be used in future work including the potential negative social impacts it could have. While the primary goal of this research is to support efforts in political research, the dataset could inadvertently be used to reinforce partisan divides. Furthermore, the speaker-focused structure of the data makes it easier to build models based on individual legislators, which raises concerns about misuse. Such models could be exploited to generate deepfakes, simulate speeches, or spread misinformation, potentially damaging public trust in political communication. These risks underscore the importance of emphasizing responsible use, transparency, and ethical guidelines for research utilizing this dataset.