From Reddit to Congressional Hearings: A Study of Representation using an Argument Extraction Method

Ju Yeon Park,[1] Arman Irani,[2] Kevin Esterling,[2,3] Michalis Faloutsos[2]

[1]The Ohio State University, [2]University of California Riverside
[3]Correspondence: kevin.esterling@ucr.edu

August 28, 2024

Abstract

In representative democracy, it is crucial to include the perspectives of those governed in policy making. To analyze representation, research often links public policy preferences with legislators' stances through surveys and votes. However, the scholarship lacks effective methods to gauge if substantive policy ideas of the public gain lawmakers' attention. This study combines Reddit discussions on policy issues with U.S. House of Representatives' hearing transcripts from 2005-2022 to develop an innovative LLM-driven argument detection and stance classification framework called WIBA ("What is Being Argued"). By applying WIBA, we visualize the overlap of arguments, identifying which communities of interest are represented or overlooked in legislative deliberations and how the pattern of representation varies across partisan and non-partisan policy issues. Our proposed approach shifts the focus from organized interests to the arguments themselves, providing a deeper understanding of democratic representation at the argument level.

1 Introduction

In representative democracy, incorporating the perspectives of those who are governed in the policymaking process is essential. Existing studies have examined the relationship between the public's policy preferences scaled through surveys, and legislators' policy stances revealed through roll call votes or elite surveys [1]. However, the scholarship currently lacks the necessary methods to examine the extent to which substantive policy ideas of the public, their arguments and reasoning, receive lawmakers' attention. This study takes the first step forward to develop methods that rely on recent Large Language Model (LLM) text analytic tools and the plethora of digital text data that have become available. Specifically, we provide a pipeline and method that evaluates the extent to which arguments that arise within communities of interest in public discourse are voiced before the U.S. Congress through committee hearings.

For our study, we link two text-based data sources – Reddit data capturing the discourse on policy issues within (self-selected) communities of interest, and the committee hearing transcripts from the U.S. House of Representatives from 2005 to 2022 (109-117th Congresses). We develop and evaluate a cutting-edge argument detection and stance classification method to assess the extent to which the ordinary discourse within communities of interest on Reddit is represented in legislative deliberation occurring in congressional committee hearings. We focus in particular on two partisan policy issues on which Americans are polarized, abortion and gun control law, and two science-based policy issues that are less polarized, GMOs and nuclear energy. The comparison across these different types of issues allows us to compare the extent to which diverse policy perspectives are considered in legislative deliberation, and whether this representation varies across partisan and nonpartisan policy issues.

Congressional committees hold hearings to collect expert information on policy issues under consideration or to learn about consequences of policy programs and their implications [3]. For this, they invite witnesses to testify in these hearings, and the witnesses include bureaucrats, researchers in universities or think tanks, representatives of specific industries, corporations, or trade associations, non-profit organizations, local governments, labor unions, etc. This practice helps Congress gather diverse viewpoints and information which impacts the decision making process when crafting legislation. A recent study provides a thorough analysis of who congressional committees tend to invite as witnesses, to indirectly learn about the flow of policy ideas and information and what they say [2]. However, in order to study representation of various perspectives that the public holds on a policy issue, we need to compare the substantive content of legislative deliberation to the substantive content of public deliberation observed among communities of interest in the general public, voiced outside of committee hearings. Thus, beyond merely assessing the composition of the organized groups represented or the identities of the witnesses themselves, our proposed methods allow us to assess the extent to which groups' perspectives are articulated by legislators or witnesses before congressional committees. With this, our novel methods introduce an innovative, argument-based approach to

study representation and quantify which social or industry groups' perspectives are better represented in the lawmaking process.

The method we propose allows evaluating the extent the voices of groups participating in the public forum are heard in committee hearings. Our analysis pipeline is driven by a holistic, argument-centric framework incorporating groundbreaking LLM methodologies for identifying the presence of arguments within a corpus, discerning the topic being argued, and classifying corresponding stances in that argument towards the topic. We call our algorithm WIBA, which stands for "What is Being Argued." The WIBA pipeline is built using advanced LLMs that have been meticulously fine-tuned for these specific tasks and rigorously evaluated against established argument mining benchmarks. Existing methods for argument mining tasks are limited by their accuracy, adaptability, and loosely defined task requirements. By introducing novel formalization for argumentation and employing innovative data augmentation techniques, these methods are agnostic to the domain the argument was made in. Moreover, our methods remain impartial to the various types of reasoning logic used in argumentation (such as deductive, inductive, abductive, etc.), enabling comprehensive coverage of the spectrum of arguments present in a corpus.

Our study demonstrates the effectiveness and robustness of these techniques whilst offering tangible insights to the role of argumentation in representation. In addition, we have created novel keyword expansion methods that we use to curate comparable corpora to make comparisons between the hearing and the Reddit data.

We use WIBA to recover the set of arguments made in each corpus. Then, we use embedding similarity to visualize the overlap in arguments between a given Reddit forum and discourse in committee hearings using a set logic displaying the intersections and complements to understand the extent to which the arguments in the Reddit forum is represented in hearings, and equally importantly, the public arguments in a forum that are excluded in hearings. By comparing these visualizations across the policy issues and across communities of interest, our methods enable us to identify which groups have a voice on a given issue and which do not. For example, we might find that the arguments made regarding GMOs within an industry-centric forum are well represented, but the arguments made within an organic vegetable gardening forum or a soil science forum are not. In sum, our methods enable us, for the first time, to conceive of democratic representation at the level of arguments, rather than at the level of organized interests, to understand whose perspectives receive consideration in committee hearings.

2 Data

2.1 Legislative Deliberation in Congressional Committee Hearings

In many democratic systems, legislative institutions such as U.S Congress or German Bundestag created committees in which a subset of lawmakers specialize in specific policy jurisdiction to divide the legislative workload. Committees are at the heart of lawmaking processes as they function as a task force. Committees hold public hearings regularly to deliberate on policy issues which may turn into legislation, study and revise bill drafts, and decide whether to propose the bills assigned to the committee for consideration of the entire chamber.[1] Thus, the transcribed text of legislative deliberation occurring in committee hearings allows us to study various policy ideas or arguments that Congress considers.

To capture the ideas and arguments used in the lawmaking process, we use U.S. House committee hearing transcript data from 2005 to 2022 (109-117th Congresses) collected from the Government Publishing Office.[2] From these original transcripts, we extracted only the statements that were spoken out loud by speakers excluding formatting texts and the texts submitted in a written form. In hearings, speakers include committee members, witnesses invited to testify in hearings, and rarely committee staff members. Typically, hearings starts with opening statements by the committee chair, ranking member and other members who chose to speak, followed by oral testimony from a panel of witnesses, which is then followed by a question and answer session where members take turns to ask questions of witnesses. When extracting and parsing the statements, we simultaneously identified the speaker identity following the procedure used in [7]. Then we merged attributes of legislators—such as their party affiliation, gender, seniority, Legislative Effectiveness Score (LES), committee chairmanship, DW-NOMINATE ideology score—from the Legislative Effectiveness Data.[3]

For the study, we collected almost 100,000 statements from 399 House committee hearings relevant to the four policy issues we consider. We selected 37 hearings on Abortion, 32 on Gun Control, 13 on GMOs, and 316 on Nuclear Energy. Table 2 presents the number of hearings, statements, and arguments in this data for each policy issue.

The selection process was deliberate: 1) We started with seed keywords and selected hearings and Reddit posts that contain these keywords, and this process was iterated independently for each of the four issues.;[4] 2) Applying BERTopic

[1] While most of the hearings serve legislative purpose, there are some other types of hearings focusing on government oversight and investigation or consideration of presidential nominations for bureaucratic posts and court justices.

[2] `https://www.govinfo.gov/app/collection/chrg`

[3] [10] for the House; [11] for the Senate. The LES measures how effective lawmaking activities a member of Congress has been engaged in during a two-year long Congress. See these works for more information.

[4] The seed keywords we used are "abortion" for Abortion, "gun control" for Gun Control, "nuclear energy" for Nuclear Energy, and "GMO", "genetically modified organism", and "genetically modified crop" for the GMOs issue.

[4] and RAKE [9] keyword expansion on the selected hearings and Reddit posts, we collected additional keywords for each policy issue. RAKE is a domain-independent keyword extraction algorithm that ranks relevant keywords based on their occurrence and co-occurence with other words. BERTopic leverages pre-trained transformer-based language model architecture to generate embeddings that are clustered together to capture coherent topics within a corpus. 3) Based on the expanded set of keywords, we repeat step 2 to finalize a set of keywords and; 4) Select the final set Reddit posts to be used for the study. For hearings, however, we took additional screening steps, we first selected the hearings in which these keywords appear for more than three times, and then we took the first 3,000 words of these hearings and asked a LLaMa 3.1 8B model with a low temperature to identify whether each hearing is primarily about the given policy issue. The prompt for this is located in the appendix.

2.2 Online Public Deliberation in Reddit

To analyze public deliberation on policy issues, we web-scraped Reddit posts relevant to the four policy issues we study. Reddit.com is one of the most popular online forum sites in the world, with 73 million daily active users and hundreds of thousands of self-selected communities called 'Subreddits'. These Subreddits are usually created around a topic of interest, ranging from extremely broad topics such as 'r/movies' which contains discussions on international films of any genre to niche topics such as 'r/chickens' which contains discussions on the care and health of chickens. Anyone can initiate public discussion by posting a thread on a Subreddit; other users can freely contribute to the discussion by posting replies to the initial thread. Table 1 presents the title and description of the eight Subreddits we consider for Abortion issue.

While online public forums are available in multiple venues, Reddit provides an ideal corpus of text data for this study for the following reasons. First, it is the largest site for online community that allows in-depth discussion on topics of users' interest. Second, by its architecture it creates self-selected communities with common interest that contribute perspectives within a channel. Hence, Reddit posts reflect policy discourse within such groups. Our methods then enable us to evaluate the extent these self-selected groups' voices are heard in congressional committees. Furthermore, it is important to note that our methods are general enough to be applied to any other text data that reflects discourse within communities of interest.

In total, we collected 32,154 Reddit threads for the analysis. Our corpus contains 6,924,263 million posts and comments from 8 Gun related Subreddits, 20,679,696 from 8 Abortion related Subreddits, 630,785 from 4 GMO related Subreddits and 457,651 from 6 Nuclear Energy related Subreddits. The data was collected from the inception of each community, as early as 2008 until February 2023. Table 2 presents the number of posts, relevant posts, and arguments for each policy issue.

These Subreddits were selected after careful consideration of their current activity status, community descriptions provided by moderators, and manual

Subreddit	Description	# Users	# Threads
r/prolife	A place for Pro-Lifers of all religious, secular, and political views to gather on Reddit.	46,000	36,206
r/Abortiondebate	Welcome to the Abortion Debate Subreddit! This Subreddit is for civil and respectful debates and discussions about abortion. All topics must be closely related to the abortion debate. Insults, ad hominems, trolling, and any other inflammatory or antagonistic language are subject to moderation and restriction of posting privileges.	9,700	9,608
r/TwoXChromosomes	Welcome to TwoXChromosomes, a Subreddit for both serious and silly content, intended for women's perspectives. We are a welcoming Subreddit and support the rights of all genders. Posts are moderated for respect, equanimity, grace, and relevance.	14,000,000	478,848
r/WomensHealth	Women's Health: women's health news, questions, and discussion. A space for women to discuss health and medicine.	115,000	40,873
r/childfree	Discussion topics and links of interest to childfree individuals."Childfree" refers to those who do not have and do not ever want children (whether biological, adopted, or otherwise).	1,500,000	264,243
r/Feminism	Welcome to the feminism community! This is a space for discussing and promoting awareness of issues related to equality for women.	291,000	85,888
r/prochoice	A sub dedicated to reproductive rights. We stand against any laws which seek to give 3rd party & government power over other people's conception and pregnancy outcomes.	45,000	23,724
r/abortion	If you're pregnant and don't want to be, we can help you get an abortion. This is a pro-abortion, stigma-free space to ask questions, get information, and share your experiences.	56,00	32,244

Table 1: Abortion Subreddits with their community descriptions and issue positions

post inspection for relevance. In order to select only the discussions and relevant information pertaining to our topic of interest, we select these posts or comments that contain *at least one* mention of one of the keywords in our expanded set.

Furthermore, for the purpose of this analysis, we reduce our number of Abortion and Gun Control related posts to increase efficiency while maintaining the diversity and representativeness of the data. Especially with our heterogeneous dataset, where different Subreddits may exhibit varying thematic and argumentation styles, traditional sampling methods such as randomized or stratified sampling may fail to capture the full breadth of diversity present in these communities.

To overcome this challenge, we introduce a Diversity-Based Sampling approach that leverages clustering techniques to ensure that the sampled data reflects the range of arguments present in the original dataset. We begin by iterating over each unique Subreddit, and define a maximum threshold of samples to select. We then apply Term Frequency-Inverse Document Frequency (TF-IDF) vectorization. Once our data has been transformed into its numerical representation, we apply the K-Means clustering algorithm to our posts,

Topic	Data Source	# Events/Posts	# Statements/Relevant Posts	# Arguments	Arg. Ratio
Abortion	Congressional Hearings	37 Hearings	9,688 Statements	5,029	52% (31.6%)
	Reddit Data	20,679,696 Posts	1,446,523 Relevant Posts	16,976*	84.88% (56.45%)
Gun Control	Congressional Hearings	32 Hearings	7,007 Statements	4,243	60.5% (33%)
	Reddit Data	6,924,263 Posts	552,033 Relevant Posts	21,376*	122% (60.11%)
GMOs	Congressional Hearings	14 Hearings	2,532 Statements	1,529	60.4% (33.24%)
	Reddit Data	630,785 Posts	5,443 Relevant Posts	5,009	92.02% (42.28%)
Nuclear Energy	Congressional Hearings	316 Hearings	66,014 Statements	36,451	55% (30.3%)
	Reddit Data	457,651 Posts	33,206 Relevant Posts	41,964	126.37% (60.40%)

Table 2: Comparison of Congressional Hearings and Reddit Data. *Sampled from a subset of relevant posts. See Section 2 for a more detailed explanation. The argument ratios for congressional hearings were measured at the statement/post level, with those in parentheses measured at the sentence level.

forming distinct clusters of similar arguments whilst also reducing any noise in our dataset. Finally, these clusters are concatenated to form the final sampled dataset.

3 Argument-centric Approach

We introduce arguments as the fundamental unit of capturing opinions, thoughts, and beliefs in this work. An argument is defined as a phrase containing at least one claim supported or attested by at least one premise. We utilize the comprehensive argument mining tool WIBA [5], which has been evaluated to perform extremely well in identifying informal and formal monological arguments made, as well as the topic being argued and associated stance of the argument. This method utilizes open-sourced LLMs that have been fine-tuned using formalized prompt-engineering and data augmentation, making it the most suitable and equitable choice for this study. WIBA is a suite of three methods, where each method takes in a text-section and outputs an argument label $\in$ {No Argument, Argument}, Topic $\in$ {Topic Argued}, and stance label $\in$ {Pro, Con}.

In order to handle long text sequences, we employ a sliding window technique which was proposed in [6]. With the size of a window or text unit set as three sentences, our algorithm determines whether the window includes an argument or not and estimates the level of confidence on this decision. Then, the window slides by one sentence. In case the windows identified as containing an argument overlaps, we select the window with the highest level of confidence.

Based on WIBA outputs, we computed the proportion of arguments conveyed in our two corpora. As we can see in Table 2, congressional hearings remained consistent overall in their argumentation across all topics, with Gun Control & GMOs being the most argumentative with roughly 60% of speeches containing an argument and 33% of sentence segments containing an argument. On Reddit, Gun Control and Nuclear Energy were the most argumentative with every post containing on average 1.22 or 1.26 arguments respectively and roughly 60% of sentence segments containing an argument.

3.1 Cross-Platform Argument Matching Framework

This section presents our framework for matching arguments made in Congressional hearings to those made on the online platform Reddit. Recognizing the distinct semantic and linguistic differences between these two platforms, our tools and methodology account for these formal and informal argument settings. Our framework is designed to systematically extract and refine, standardize, and match arguments irrespective of their source, enabling the quantification of intensity of representation of Reddit discourse within congressional hearings.

Argument Extraction and Representation. We now propose a pipeline for the distillation of argumentation across vastly different platforms. We acknowledge that argument styles and semantics differ across traditional oratorical arguments than those digitally and informally expressed. One key assumption is that WIBA is able to identify monological arguments in these environments with high accuracy. For arguments resembling those that are found in informal, social media settings WIBA achieves an accuracy of 76%. For arguments made in legislative environments, the accuracy is 89% [5]. Given this, we proceed to the next step of our representation process. We have every argument combined with their topic and stance information, e.g., "Argument;Topic;Stance" to encode even more meaning, and embed these using SentenceTransformers all-mpnet-v2 model [8], into a vector space of 768 dimensions, which has shown tremendous ability in capturing semantic meaning and information.

Argument Standardization. Once we have our argument information transformed into a vector embedding, we choose to normalize the embeddings using a StandardScaler which scales the embeddings to have a mean of 0 and a standard deviation of 1. This process is employed to ensure quality, accuracy, and consistency of later downstream clustering tasks.

Matching Reddit Arguments to Congressional Arguments. In order to investigate the representation of Subreddit community's discourse within U.S congressional hearings, we developed a computational method to match those arguments made in Reddit communities to those made in Congressional hearings.

For each unique Subreddit, the arguments that were made are isolated, and their embeddings were compared against all congressional arguments. This comparison took each Reddit argument and calculated its cosine similarity to each of the congressional arguments embedding. Those above a similarity threshold of 0.7 were kept and considered to be similar. The threshold of 0.7 was selected to focus on significant similarity, whilst also balancing the strictness of the comparison. We select cosine similarity as our similarity measure as it is commonly used in NLP tasks to assess the similarity between two vector embeddings.

Through manual investigation, we see tremendous success of our methods in matching arguments made in Reddit to those made in congressional hearings. Here we see a randomly selected example of a matched congressional argument to an argument made in Reddit. Both arguments are skeptical about the effect of abstinence education in reducing the rate of teenage sexual activity.

Congressional Argument

"You think, if we're increasing the number of teens who are abstaining, won't that automatically reduce teen pregnancy. Not necessarily. My concern is that abstinence-only programs may actually decrease contraceptive use among teens who ultimately decide to have intercourse."

Reddit Argument

"The rate of teenage sexual activity has decreased by about 25% within the last generation, which almost completely explains the decrease in the teen birth rate. I doubt abstinence education had a lot to do with it, but obviously contraception availability does not decrease sexual activity."

3.2 Quantifying Representation

Based on the arguments matched across two platforms, we propose two novel metrics of representation. The first metric is Representation Intensity, $\mathcal{RI}(\mathcal{S})$, which quantifies the extent to which the arguments made in a public community receives attention in legislative deliberation. To that end, we compute the percentage of Subreddit arguments represented in congressional Hearings. That is, we compute the ratio of arguments in a Subreddit community that were matched to those in congressional hearings to the total number of arguments within the Subreddit, and we multiply 100 to turn it into percentages. Ultimately, this metric provides insights into the dominance or marginalization of laymen discourse in political environments.

The Representation Intensity $\mathcal{RI}(\mathcal{S})$ for a Subreddit $\mathcal{S}$ within congressional hearings C is defined as:

$$\mathcal{RI}(\mathcal{S}) = \frac{|A_{\mathcal{S}}(C)|}{|A(S)|}$$

Where $|A_s(C)|$ is the number of arguments from a Subreddit $\mathcal{S}$ within the set of Congressional Arguments C, and $|A(\mathcal{S})|$ represents the cardinality of the entire set of arguments made in that Subreddit. A higher value of $\mathcal{RI}(\mathcal{S})$ indicates a more intense representation of a Reddit Community $\mathcal{S}$, while a lower value suggests a more sparse or diluted representation of specific arguments in legislative discourse.

The second measure is the Representation Intensity $\mathcal{RI}(\mathcal{L})$ of individual legislators which captures the extent to which the arguments a legislator makes during committee hearings tend to align with the arguments made in Reddit on a given policy issue. For this, we compute the ratio of a legislator's arguments that were matched to those in Reddit to the total number of arguments that they made in the hearings on a given policy issue. Formally, the Representation Intensity $\mathcal{RI}(L)$ for a legislator L within congressional hearings is defined as:

$$\mathcal{RI}(\mathcal{L}) = \frac{|A_{\mathcal{L}}(S)|}{|A(L)|}$$

Where $|A_{\mathcal{L}}(S)|$ is the number of arguments that a legislator $\mathcal{L}$ made within the set of Subreddit Arguments S, and $|A(L)|$ is the total number of arguments that $\mathcal{L}$ made in hearings on a given policy issue across the period that we analyze.

(Similarly, it is possible to construct a measure for witnesses to understand which types of witnesses tend to present arguments aligned with public discourse in their hearing testimonies.)

3.3 Analysis

3.3.1 Representation of Subreddit Communities

To provide a visual representation of the degree of $\mathcal{RI}(\mathcal{S})$ for each Subreddit, we first calculate the percentage of arguments in each Subreddit that meets or passes the threshold of the similarity to congressional arguments. These percentages are then visualized using a bar chart, where the height of the bars represent the $\mathcal{RI}(\mathcal{S})$ of the Subreddit.

A considerable amount of arguments made on Reddit are not represented in Congress. This methods provides nuanced insight into how representative specific community's arguments are represented within congressional Hearings. Overall, we find Reddit arguments having a $\mathcal{RI}(\mathcal{S})$ of 39.95% in Congress. The per Subreddit $\mathcal{RI}(\mathcal{S})$ can be seen in Figure 1. The arguments in "Abortiondebate" Subreddit are most likely to be mentioned in congressional hearings among all. Given that this Subreddit is described to invite discussion from both pro- and anti-abortion sides, this result makes sense. The second best represented Subreddit is "prochoice" followed by "prolife", which together constitute the major perspectives discussed on abortion. Then, around 4-5% of the arguments made in the communities advocating gender rights, namely, "Feminism" and "TwoXChromosomes," are represented. Given the size of the community and text corpus of the "abortion" Subreddit, the $\mathcal{RI}(\mathcal{S})$ is surprisingly low, which is probably because the extremity of the discussions exchanged in this community as illustrated in its description in Table 1. The least represented community being 'r/womenshealth' which is a space for women's health news, questions, and discussion has a $\mathcal{RI}(\mathcal{S})$ of 0.3%.

3.3.2 Legislators' Representation Efforts

Which legislators tend to reflect public discourse? We also analyze how representative individual legislators are of laypeople's beliefs. The total number of arguments that members of Congress made on the abortion issue ranges from 1 to 107 with its mean at 6.82. The number of arguments matched to all eight Subreddits on abortion ranges from 0 to 71, and the overall match rate, $\mathcal{RI}(\mathcal{L})$, ranges from 0 to 100 with its mean at 54%.

To examine which legislators' arguments tend to be aligned with those appearing Reddit, we conducted an OLS regression analysis where the unit of analysis is individual legislator and the dependent variable is the number of

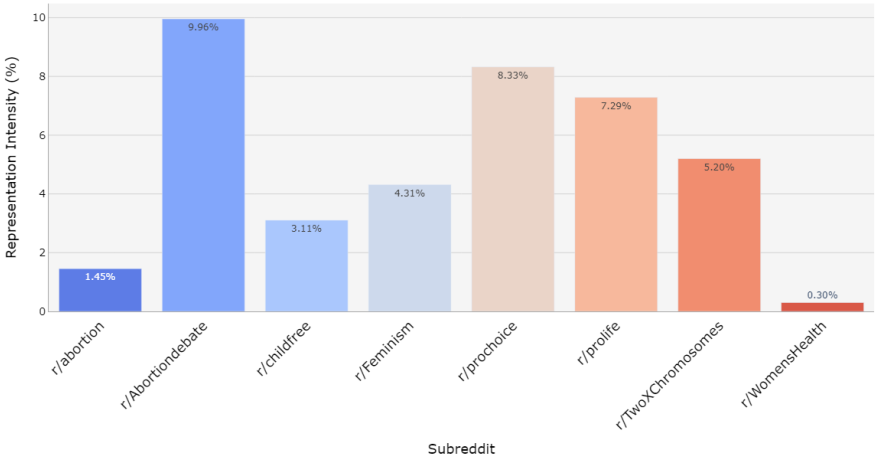

Figure 1: The Percentage of Subreddit Arguments Mentioned in Congressional Hearings.

their arguments matched to those in each Subreddit. Independent variables include the indicators for Democrats, females, and committee chairs, Legislative Effectiveness Score, and the number of terms they served in the chamber.[5] Lastly, we control for the total number of arguments they made on a given policy issue. In addition to the models fit on each of the eight Subreddits, we also fit a model for all Subreddits with the total number of arguments that were matched to any of the Subreddits as a dependent variable. The results are shown in Table 3.

Democrats and ideological extremists tend to better represent Reddit users. As shown in the first model, in general, Democratic Party members and those with extreme ideological positions tend to make arguments that were mentioned in Reddit.[6]

Next, we present results for the eight Subreddits in a meaningful order. Based on the description of each Subreddit, we ordered them from anti-abortion to pro-abortion, and the order is reflected in in Table 1.[7] Thus, in Table 3, the

[5] As we aggregated the data by legislator across all congresses they served, except for the partisanship and gender, the average values of time-varying variables were used. We could fit models with individual fixed effects using the panel data, but this approach would fail to measure the effects of partisanship and gender of legislators which are expected to be most relevant predictors on the abortion issue.

[6] Ideological Extremism variable is the absolute value of the DW-NOMINATE score.

[7] The reasoning for the order and our ad hoc issue position scores (-2 to 2) are as follows: 1) "prolife" is the only community representing an anti-abortion perspective (-1.5); 2) "Abortiondebate" takes a neutral position inviting discussions from both pro- and anti-abortion sides (0); 3) "TwoXChromosomes" and "WomensHealth" are leaning towards pro-abortion, but these communities are not primarily about abortion (0.5); 4) "childfree" and "Feminism"

models to the right are for more explicitly pro-abortion Subreddits and those to the left are less so. The results remain largely consistent with the first model including all Subreddits. However, there are three notable points. First, Democrats are found to better represent the arguments of the communities more strongly leaning towards pro-abortion communities than those less so. Second, female legislators tend to better reflect arguments in communities representing women's benefits "WomensHealth". Third, ideologically extreme legislators tend to speak in alignment with the communities that are anti-abortion (e.g. "prolife") or weakly related to pro-abortion perspectives (e.g. "Abortiondebate" and "TwoXChromosomes").

4 Visualization of Argument Clusters

In this section, we outline our comprehensive computational framework designed to transform raw arguments from both congressional hearings and Reddit discussions into structured and meaningful insights. By leveraging the state-of-the-art techniques outlined, we unify a diverse corpus of monological arguments to facilitate the extraction and analysis of core themes across formal legislative settings and informal online forums.

Rather than relying on our predefined Subreddit groups, we take an unsupervised clustering methods through which the natural structure of arguments to emerge. This approach enables us to capture the nuanced distinctions between overall themes present in the dataset to better understand the distribution and emphasis of discourse in these two different environments.

4.1 Argument Refinement: Extracting Core Themes from Complex Discourses

Dimensionality Reduction. We employ the widely used Principal Component Analysis (PCA) technique, which reduces the number of dimensions while retaining the most important features. Particularly for our dataset of congressional and Reddit arguments, which can be noisy and sparse, this preprocessing step makes subsequent clustering tasks more effective.

Preserving Information. Uniform Manifold Approximation and Projection (UMAP), a non-linear dimensionality reduction technique, is next applied. This technique is effective as preserving both the global and local structure of our data, which in the context of this analysis is the cross-corpora arguments and local arguments made within each platform. This preservation is ideal for capturing the nuanced relationships between arguments from different origins, such as congressional hearings and Reddit. Furthermore, reducing our data's dimensions further, to two dimensions, allows for the creation of visually inter-

Subreddits are not primarily abortion but closer to this issue by their nature (1); 5) "pro-choice" is the intuitive opposite of "pro-life" Subreddit 91.5); 6) "abortion" takes the most extreme position suggesting to help users get an abortion (2).

Table 3: Legislators' Arguments Matched to Reddit Arguments

	Overall	prolife	Abortionde...	TwoXCh...	WomensHealth	childfree	Feminism	prochoice	abortion
Democrat	0.790**	0.394	0.400	0.752**	0.090	0.884***	1.023***	0.692**	0.322
	(0.315)	(0.297)	(0.300)	(0.290)	(0.124)	(0.318)	(0.378)	(0.323)	(0.212)
Female	0.401	0.111	0.351	0.520*	0.257**	0.479	0.509	0.348	0.234
	(0.302)	(0.285)	(0.287)	(0.278)	(0.118)	(0.304)	(0.362)	(0.309)	(0.203)
LES	-0.052	-0.017	-0.024	-0.053	0.026	-0.037	-0.033	-0.068	-0.069
	(0.093)	(0.088)	(0.089)	(0.086)	(0.037)	(0.094)	(0.112)	(0.096)	(0.063)
Ideological Extremism	2.126**	1.857**	1.632*	1.962**	0.384	1.075	1.701	1.272	-0.115
	(0.962)	(0.906)	(0.915)	(0.884)	(0.377)	(0.969)	(1.153)	(0.985)	(0.647)
Seniority	-0.021	-0.056*	0.003	0.018	0.011	-0.027	-0.006	-0.013	0.0004
	(0.033)	(0.031)	(0.031)	(0.030)	(0.013)	(0.033)	(0.039)	(0.034)	(0.022)
Committee Chair	1.184*	0.540	0.770	0.971	0.260	0.714	1.208	1.103	0.104
	(0.709)	(0.668)	(0.674)	(0.651)	(0.278)	(0.714)	(0.850)	(0.726)	(0.477)
Num. Arguments	0.633***	0.376***	0.400***	0.384***	0.038***	0.332***	0.422***	0.427***	0.159***
	(0.012)	(0.011)	(0.011)	(0.011)	(0.005)	(0.012)	(0.014)	(0.012)	(0.008)
Constant	-1.843***	-1.149**	-1.402**	-1.769***	-0.395*	-1.383**	-2.017***	-1.264**	-0.417
	(0.599)	(0.565)	(0.570)	(0.551)	(0.235)	(0.604)	(0.719)	(0.614)	(0.403)
Observations	166	166	166	166	166	166	166	166	166
R^2	0.950	0.884	0.895	0.895	0.353	0.840	0.858	0.893	0.725
Adjusted R^2	0.948	0.878	0.890	0.890	0.324	0.833	0.852	0.888	0.713

Note: *p<0.1; **p<0.05; ***p<0.01

pretable plots that reveal clusters and patterns that may otherwise be hidden in higher dimensions.

Probabilistic Clustering. Once we have this new lower-dimensional representation of our data, we continue clustering to identify distinct yet coherent communities of narratives expressed. We employ the use of Gaussian Mixture Models (GMM), which is a probabilistic clustering technique that models the data as a mixture of multiple Gaussian distributions. Each Gaussian corresponds to a cluster, with the model assigning probabilities to each argument belonging to a particular cluster.

The motivation for this clustering technique is twofold. For one, unlike traditional clustering techniques such as K-means or HDBSCAN, GMM has the flexibility of modeling clusters that are not necessary spherical. This is crucial for dealing with our argument embeddings, as clusters may have various shapes and orientations due to its diversity and complexity. Furthermore, GMM is a soft clustering technique which assigns a probability of belonging to multiple clusters, which is particularly useful when arguments may be similar across different topics and stances.

Cluster Comprehension using Large Language Models. As part of the comprehension portion of being able to decipher and extract meaning from the clusters that we create, we propose a multistep approach to generating argument themes. We employ the use of LLaMa 3.1 8B, to iteratively summarize and refine the core concepts of arguments found in a cluster.

Since each cluster may contain thousands of arguments and therefore thousands of sentences, we tackle the first problem of context length and comprehension of LLMs by summarizing batches of arguments at a time. Each cluster has 50 of its arguments summarized at a time, and then these summaries are saved to be passed into a final summarization process. The LLMs temperature is set very low (0.1) and the prompt design is located in the appendix for reproducibility.

4.2 Creating Interpretable Visualizations

Visualizing Political Discourse on Cross-platforms. Figure 2 is a projection into two-dimensional space, the deliberation surrounding the topic Abortion. Each point represents a unique argument made on either Reddit (red dot) or in Congress (blue dot). The proximity of two dots to each other represents the semantic similarity of these arguments, which we visually cluster together using circles to encapsulate clusters or groups of similar arguments. The cluster summaries are then provided as annotations into the visualization, providing an informative and intuitive understanding of how different arguments are distributed in the reduced dimensional space. This enables researchers to quickly grasp the primary factors influencing the distribution and overlap of representation in the projection.

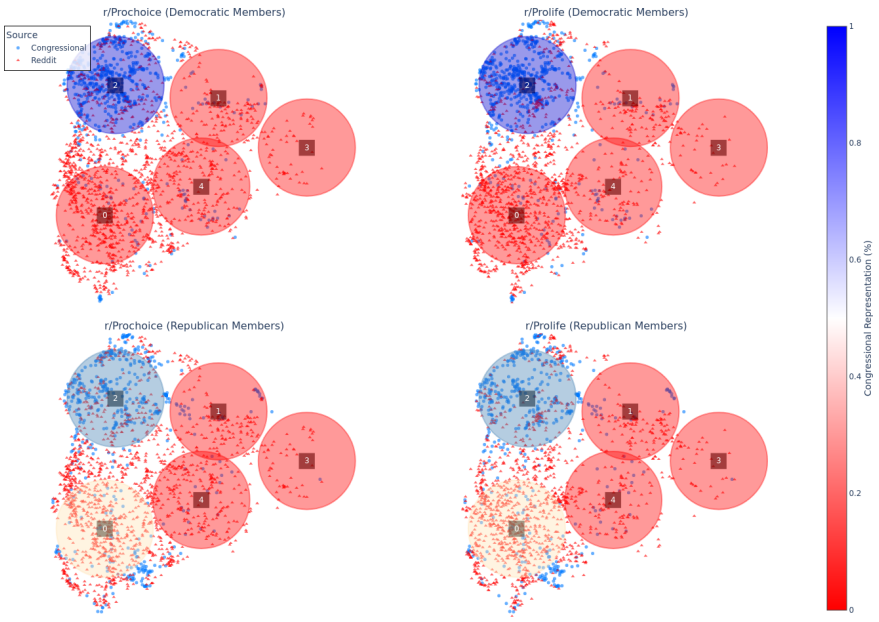

Figure 2: 2D projection of arguments made in Congressional Hearings and on Reddit forums using our Argument-driven approach. Each cluster is filled with a color representing its Congressional Representation Intensity. Darker blue indicates more Congressional Representation and more Red represents less Congressional Representation. Each cluster is labeled with a number: (0) Abortion Rights (1) Sex Education and Birth Control (2) Pros and Cons on Abortion (3) Birth Control (4) Women's Well-being and Bodily Autonomy.

4.3 Results

Modeling Representation Shifts in Congressional and Reddit Argument Clusters. We conduct an analysis investigating the composition of the clusters that were generated based on our method, in order to model the change in representation based on controlled variables. Our approach models the entire Reddit and Congress argument corpora and then filters the data based on two variables: 1) the Subreddit 'r/prolife' and 'r/prochoice', two communities with polar opposite views, and 2) legislators' party affiliation of either Democrats or Republicans. By controlling for these variables, we observe the resulting representation shifts in our identified argument clusters.

The analysis reveals that Cluster # 2 (Pros and Cons on Abortion) has a significant increase in congressional representation, roughly 20% with Democrat members. Conversely, Cluster # 0 (Abortion Rights) exhibits a 15% increase in Representation with Republican members. These findings offer an initial insight into how well Congress is representing the views of distinct communities, and are a preliminary step into more detailed and nuanced findings.

5 Conclusion

In this paper, we demonstrate a new text analytic pipeline to study *argument-centric representation* by comparing policy-relevant arguments as they occur in an online public forum and legislative hearings. Specifically, we introduce the WIBA framework and apply it to Reddit posts and U.S. congressional committee hearing transcripts on (four) policy issues from 2005 to 2022. We extracted arguments from both platforms and computed the extent to which the policy ideas or arguments of each public community are articulated in the committee-stage lawmaking process. Reddit is an ideal text data source for this comparison because its forums consist of self-selected communities of interest and the discussion threads hosted in each forum enable us to observe naturalistic, policy-related discourse that occurs in diverse communities of interest.

We use WIBA to evaluate the extent to which each community's arguments are heard within the formal debates before congressional committees, and assess which communities have their perspectives represented and which do not. We find that the percentage of arguments in public discourse that are articulated in legislative process tends to be low in general. Nonetheless, our results demonstrate variations in the Representation Intensity across different Subreddits such that those representing mainstream ideas on the issue are better represented while extreme communities receive less attention in legislative deliberation.

In addition, we present the measure of representation for individual legislators and find that Democrats, ideological extremists and female legislators are more likely to put forward arguments that are similar to those of the Reddit users. Given that our current analysis focuses on abortion, these findings might be specific to this issue. Additional analysis on various policy issues may help revealing a more general pattern of argument representation.

The novel methodological approach we present in this paper makes an innovative contribution to the study of democratic representation. It showcases for the first time the possibility of analyzing the prevalence of arguments across forums through a relatively automated process at a large scale using text corpora but at a more granular level. This approach provides new insights into which communities of interest has its perspectives heard before congressional committees. Further, it can be used to identify which legislators and which interest groups articulate those perspectives. This approach augments previous work that considers the identities of the interest groups and witnesses that are called to testify. Instead of focusing on the descriptive representation of groups, we are able to focus on the substantive representation of communities of interest, that is, to identify which communities of interest have a voice in Congress beyond those that happen to be invited to testify.

Here we only offer the initial descriptives in our comparisons. In future work, we plan to model the determinants of the prevalence of different types of arguments before committees. Moreover, the methods are general and can be used to make comparisons in argumentation across any text datasets. For example, the reports from interest groups or think tanks can be used in place of the Reddit data to analyze how the arguments from these groups are reflected in legislative deliberation. Beyond the representation, the argument extraction and comparison techniques can be used to study the flow of ideas between two contexts of any kind.

6 Appendix

6.0.1 Prompt for Hearing Relevancy

```
"This is an excerpt from a US Congressional committee hearing
transcript.  Can you please tell me whether this hearing is
primarily about abortion/gun control law/nuclear
energy/Genetically Modified Organisms (GMOs)?  Please answer
either yes or no."
```

6.0.2 Prompt for Cluster Summarization

```
You are an AI expert in analyzing and synthesizing arguments from
diverse sources, including congressional hearings and Reddit
discussions.  Your task is to:

1. Identify the core themes and main points within the provided
   arguments or set of arguments.

2. Synthesize these themes into a concise summary, focusing on
   the key ideas expressed.

3. Express this summary in a clear and coherent paragraph of
   2-3 sentences.
```

```
Guidelines:

- Capture the main themes and recurring points without losing
  important details.
- Maintain objectivity and impartiality in your analysis.
- Ensure that the summary is comprehensive yet concise,
  covering the core ideas without unnecessary elaboration.
- Do not output anything else besides the summary.  No, "here
  is the summary" or similar phrases.  Output the summary as a
  brief paragraph (2-3 sentences) that encapsulates the main
  ideas.
```

References

[1] Bafumi, J., Herron, M.C.: Leapfrog representation and extremism: A study of american voters and their members in congress. American Political Science Review **104**(3), 519–542 (2010)

[2] Ban, P., Park, J.Y., You, H.Y.: How are politicians informed? witnesses and information provision in congress. American Political Science Review **117**(1), 122–139 (2023)

[3] Esterling, K.: Buying expertise: Campaign contributions and attention to policy analysis in congressional committees. American Political Science Review **101**(1), 93–109 (2007)

[4] Grootendorst, M.: Bertopic: Neural topic modeling with a class-based tf-idf procedure. arXiv preprint arXiv:2203.05794 (2022)

[5] Irani, A., Park, J.Y., Esterling, K., Faloutsos, M.: Wiba: What is being argued? a comprehensive approach to argument mining. arXiv preprint arXiv:2405.00828 (2024)

[6] Irani, A., Park, J.Y., Esterling, K., Faloutsos, M.: Wibacong: An argument-centric framework for understanding us congressional hearings. arXiv preprint arXiv:2407.06149 (2024)

[7] Park, J.Y.: When do politicians grandstand? measuring message politics in committee hearings. Journal of Politics **83**(1), 214–228 (2021)

[8] Reimers, N., Gurevych, I.: Sentence-bert: Sentence embeddings using siamese bert-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics (11 2019), https://arxiv.org/abs/1908.10084

[9] Rose, S., Engel, D., Cramer, N., Cowley, W.: Automatic Keyword Extraction from Individual Documents, pp. 1 – 20 (03 2010). https://doi.org/10.1002/9780470689646.ch1

[10] Volden, C., Wiseman, A.E.: Legislative Effectiveness in the United States Congress: The Lawmakers. Cambridge University Press, New York (2014)

[11] Volden, C., Wiseman, A.E.: Legislative effectiveness in the united states senate. Journal of Politics **80**(2), 731–735 (2018)