What Is Being Argued (WIBA)? An Application to Legislative Deliberation in the U.S. Congress

Arman Irani¹, Ju Yeon Park², Kevin Esterling¹, Michalis Faloutsos¹

¹UC Riverside, ²The Ohio State University

Abstract

How can we utilize state-of-the-art NLP tools to better understand legislative deliberation? Committee hearings are a core feature of any legislature, and they offer an institutional setting which promotes the exchange of arguments and reasoning that directly impact and shape legislation. We apply What Is Being Argued (WIBA), which is an argument extraction and analysis framework that we previously developed, to U.S. Congressional committee hearings from 2005 to 2023 (109th to 117th Congresses). Then, we further expand WIBA by introducing new ways to quantify various dynamics of democratic deliberation. Specifically, these extensions present a variety of summary statistics capturing how deliberative or controversial a discourse was, as well as useful visualizations to the WIBA output that aid analyzing arguments made during the legislative deliberation. Our application reveals potential biases in the committee system, and how political parties control the flow of information in 'hot topic' hearings.

1 Introduction

This paper demonstrates the application of What Is Being Argued (WIBA), an argument-centric NLP framework that we have previously developed,¹ to political texts – in particular policy deliberations occurring in U.S. congressional committee hearings. Additionally, this paper further extends WIBA by proposing useful visualization and summary metrics using WIBA output to quantify the nature of the discourse. Thus, here we demonstrate the power of WIBA to unpack and analyze political discourse in this legislative deliberation setting.



Figure 1: The dashboard for viewing hearing level deliberation dynamics. Each bar represents an argument presented by a speaker, and each unique color represents a cluster the argument belongs to. The horizontal axis labels last names of the legislators with their partisan affiliation in parenthesis: "D" for Democrats and "R" for Republicans. Witnesses are marked with "W" followed by a number depending on the order of their first statement.

Problem: Political proceedings often capture dialogue in a transcript, and so generate an immense amount of text that can be mined in order to better understand deliberative democracies. The focus of this paper is specifically on U.S. congressional committee hearings, which are central to policymaking in the U.S. Congress. Although these proceedings are incredibly important for the function of Congress, there has been little progress in using computational text analysis methods to analyze them for their deliberative quality. Thousands of hearings take place every year in the United States, and it is impossible to make sense of such a large amount of information in its raw form.

Solution: We propose to use WIBA, an argument-centric approach we developed, to understand arguments and reasoning exchanged in legislative deliberation occurring during the U.S. congressional committee hearings. Formally, we

¹To see the underlying paper and to learn the methodological details for WIBA, please navigate to https://arxiv.org/abs/2405.00828. To see the accompanying interactive website where users can try out WIBA on their own text submissions, please visit https://wiba.dev/. The project Github is located at https://github.com/Armaniii/WIBA.

define an *argument* as a statement where a claim is supported by at least one premise. Given any corpus, WIBA completes three tasks: It identifies arguments, the topic being argued, and the stances of these arguments for or against the topic. As a result, we are able to deeply understand the narratives expressed, the diversity of these narratives, and how these ideas flow and interact with one another.

Contribution: This paper proposes various quantification and visualization strategies for the WIBA output in order to better understand the deliberation and debate dynamics of U.S. committee hearings. First, we quantify argumentativeness at a statement, speaker, and hearing levels by computing the proportion of arguments among the sentences conveyed in the text. Second, WIBA uses an argument similarity model to cluster similar argument themes together. Using these two features, in this paper, we develop a pipeline to calculate the Deliberation Intensity Score of any given unit of texts (e.g., speakers or hearings) and to visualize and quantify the exchange of ideas or arguments in a conversation among multiple speakers (e.g., hearings). These methods and framework developed are a novel NLP-driven approach for gaining a deeper understanding into the democratic processes of governments. While this paper adapts WIBA to the specific context of U.S. congressional hearings, the insights from this framework can be utilized in any institutional setting where discourse has been transcribed to text.

2 Methodology

This section outlines the computational elements that we newly introduce in this paper as an extension of WIBA, and below we apply these elements to a case study.

2.1 Argument Mining

In a previous paper, we developed WIBA (Irani et al., 2024b), a systematic approach to enable the comprehensive understanding of What Is Being Argued. At a high level, our approach leverages the fine-tuning of Large Language Models along with prompt engineering to identify (1) The existence of an argument (2) The topic being argued and (3) the argument stance towards the topic. An argument is defined as a statement that contains at least one claim that is supported by at least one premise. In this previous paper we demonstrated the high performance capability of WIBA, especially in its ability to handle both informal (e.g., Reddit or Twitter) and formal (e.g, legal or political proceedings) types of arguments.

Our methods identify arguments and their contents, but do not make an assessment of the validity or truth of the arguments. Such an assessment is not necessary for our methodological purposes, nor is it normatively necessary; for example, democracies are designed to enable true and false arguments to compete rather than to have some third party determine their validity (Jefferson, 1823).

2.2 Legislative Argument Detection

To address the challenge of identifying arguments in long text sections where the exact span of an argument is unknown, we employ a *sliding window* technique. For each statement or utterance made by a speaker, we consider every possible three consecutive sentences and call it a text unit to be analyzed. This window size of three was chosen since the argument detection model we use, WIBA-Detect, was trained on a dataset with an average text unit size of three sentences. We define the step size to be one, so the text unit will move forward by one sentence.

Our argument detection model, WIBA-Detect, analyzes each text unit, determines whether those three sentences in the text unit forms an argument or not, and assigns a binary label to the text unit: {NoArgumnet, Argument}. Along with the label, it computes a confidence value for the decision, and those receiving over 0.5 level of confidence are considered an argument. Since the step size is one, there will be overlapping windows with varying confidence scores, as shown in Figure 2. If two overlapping windows are both labeled as an argument, the window that has the higher confidence can keep its "Argument" label initially assigned, but the label for the other window is adjusted to be "NoArgument." This ensures that only the most obvious argumentative text unit is labeled as an argument, but not the text surrounding it.

We repeat this process for both Argumentative Topic Extraction (WIBA-Extract), and Argument Stance Detection (WIBA-Stance).

2.2.1 Speaker Argumentativeness

Political theories may require testing which person tends to provide more arguments in their statements. We propose a method for calculating the *argumentativeness* of a particular speaker, who is active in



Figure 2: An example of our automated sliding window Argument Detection process. In this example, the second window has a higher argument confidence, therefore we assume this text unit is more of an argument than the first window.

congressional hearings.

First, all of a speaker's statements are collected with the following information per statement: 1) The number of sentences in statement, ignoring introductory and filler sentences, e.g., "Thank you Mr. X." or "Okay," etc. 2) The number of arguments made within a statement. Since each argument unit consists of a constant length of three sentences, we multiply the number of arguments by three in order to get an accurate representation of argumentative composition.

For a given speaker, the argumentativeness is defined by the number of arguments made by the speaker divided by the total number of sentences that the speaker spoke in the hearing.

$$ARG_{speaker} = \frac{3 * \# \text{ Arguments}}{\# \text{ Sentences by Speaker}}$$
(1)

2.3 Hearing Argumentativeness

Some political discourses may convey more argumentation than others. For example, hearings on contentious issues may draw more arguments or reasoning to convince others than simple exchange of numerical reports. To enable such analysis, we propose a formula for calculating the argumentativeness of a given hearing, which we define as the count of arguments present in a hearing, multiplied by three, the text unit size, divided by the total number of sentences in the hearing.

$$ARG_h = \frac{3 * \# \text{Arguments}}{\# \text{Sentences in Hearing}}$$
(2)

2.4 Thematic Argument Similarity

A critical component to the analyses of this work, is the ability to measure the degree of similarity between arguments. To do so, we utilize the state-of-the-art Sentence Transformer model, 'all-mpnet-v2', which has been fine-tuned on 1 billion sentence pairs using contrastive training. This model calculates the similarity between two sentences or paragraphs (up to 384 words) on a scale of 0 to 1. A score of 0 means the two texts are completely dissimilar, and a score of 1 means the two texts are identical.

Testing on the Best-Worst Scaling (BWS) Argument Similarity corpus provided by Ubiquitous Knowledge Processing Lab (UKP), we obtain a Cosine F_1 score of 70.2%. This result indicates a superior ability to determine the narrative similarities between arguments, as the BWS corpus consists of 3,400 pro/con stance arguments across eight controversial topics (Thakur et al., 2021).

2.5 Quantifying Deliberation Intensity

How diverse arguments or ideas are exchanged in a political discourse? We define the idea of *Deliberation Intensity* as a metric to quantify the amount of deliberation taking place, at both a speaker and hearing levels. We are not suggesting that the Deliberation Intensity is a proxy of deliberation quality (such as what is measured in the Discourse Quality Index), but rather our measure of deliberative intensity is a measure of the variety of argumentation made in a discourse (Irani et al., 2024a).

We propose to measure Deliberation Intensity of a statement using the following formula, modified from the aforementioned paper. 2

$$D_{Cluster} = \frac{\# \text{Clusters}}{\# \text{Arguments}}$$
(3)

To count the number of clusters, we first group all arguments that are similar to each other into clusters and then count the number of these clusters. Intuitively, cluster diversity captures the variation of arguments within a hearing as the percentage of arguments that are unique.

We now define our Deliberation Intensity Score (DIS) of a hearing as follows:

$$DIS = \sigma_1 * D_{Cluster} + \sigma_2 * ARG_h \qquad (4)$$

²The original formulation was designed for discussions with nested structures, like those found on Reddit.com.



Figure 3: Interactive dashboard for viewing Hearing-level deliberation dynamics.

Setting the weights σ_1 and σ_2 on the terms comprising the score allows us to put more emphasis on the diversity of interest. Here, we define their values with the use of logit functions: $a_1 = \frac{1}{1+e^{-(\# \text{Arguments})}}$ and $a_2 = \frac{1}{\frac{1+e^{-(\# \text{Total Statements})}}$. We set $\sigma_1 = \frac{a_1}{a_1+a_2}$ and $\sigma_2 = \frac{a_2}{a_1+a_2}$. The logit functions use the data to assign weights, with the intuition if the denominator of the ratio increases, the importance increases. For example, 30% of unique arguments out of 20 arguments may carry more value to the diversity of arguments than the same ratio out of 5 arguments.

2.6 Controversiality Measurement

As a representation of controversiality, we calculate the difference in the count of pro/con stances for either a given topic, or hearing. The stances are generated using WIBA-Stance, and the difference as the ratio of the smaller number of arguments with a given stance to the larger number of arguments with a given stance, multiplied by 100 to obtain a percentage. This percentage is referred to as the Controversiality Score, with the following formulation: $\left(\frac{\text{Smaller Stance Number of Args}}{\text{Larger Stance Number of Args}} \times 100\right)$

2.7 Visualizing Deliberation Dynamics

The highlight of this paper is the methodology we propose to create a visually informative and interactive figure as shown in Figure 1. For a given hearing, we process the statements first by filtering by arguments presented. Once the data consists of arguments presented by each speaker in chronological order, a community detection algorithm is run to generate clusters of arguments based on their semantic similarity. Each cluster represents a group of arguments that share common themes or topics. Each cluster is summarized utilizing LLaMa 3-8B and few-shot prompting to generate a concise summary of the key points being argued. Furthermore, speakers are categorized and labeled based on their roles (member or witness) and party affiliation. A horizontal bar plot is created to represent the duration and timing of each argument, color-coded by cluster.

This function aims to aid researchers in understanding the flow of changing themes in political discourse. Using our dashboard, as shown in Figure 3, researchers are able to generate a deliberation interaction graph for a selected hearing. The chart will populate with member and witness information, which can be observed if hovering over any dialogue box. Members information includes the speaker's political party, if their party was in the majority, the state they are representing and a distilled key point summary of the argument presented in each statement they make in a hearing. A witness's information consists of their affiliation as well as the key point summary of the argument presented in each statement they make. Each box in the chart represents an argument made by the speaker, and a color is assigned to the box depending on which cluster the argument belongs to. A gray cluster represents no assignment. A legend is included with a cluster's most frequent argued topics, which can be isolated by selection to reveal in the chart only arguments about that topic. It is important to note that there may be duplicate topics in the legend 3 ,

³These topics are automatically generated by concatenating the most frequent WIBA-Extract topics present in the

represented by different colors, which indicates two arguments being made about the same topic, but with a *different* narrative. Finally, all clusters have a concise summary contained within easy to read boxes below the chart, which provides researchers with an at-glance understanding of all the different arguments made in the selected hearing.

3 Case Study

3.1 Data

We test our methods upon a collection of U.S. congressional committee hearings on abortion and Genetically Modified Organisms (GMOs). To identify hearings on each of these particular policy issues, we started with a seed keyword (e.g., abortion and Genetically Modified Organisms or GMOs) and selected all the hearings that contain these seed words three times or more. Then, we used BERTbased keyword extraction tool (BERT-Topic) as well as the RAKE Python package designed for keyword extension, and treated each statement in a hearing as a document to identify topics commonly talked about in order to expand our keyword set. Using the expanded set of keywords for each policy issue, we counted the number of appearances of these keywords in each hearing. Then, we selected hearings in which these keywords together appeared for more than ten times. We asked Chat-GPT 4.0 whether each hearing we pre-selected is primarily about the given policy issue based on the first 3,000 words spoken in the hearing. Based on the results from these queries, we eventually identified 14 hearings on GMOs and 26 on abortion. There are 63 unique witnesses and 133 members for GMO hearings, with a total of 2,810 statements made. For Abortion hearings there are 109 unique witnesses, and 193 members with a total of 7,363 statements made.

3.2 Findings

This section presents interesting patterns observed in these hearings using WIBA output and new metrics we introduce in this study.

Abortion hearings have higher Deliberation Intensity than GMO hearings. By calculating the DIS for hearings on GMOs versus hearings on Abortion, we find a stark difference in the intensity, which can be seen in Figure 4. We additionally performed a t-test and determined the difference was statistically significant, with a p-value of 0.007.



Figure 4: The Empirical Cumulative Distribution Function (ECDF) Plot for Deliberation Intensity Score for US Congressional Hearings on Abortion & GMOs.

The result suggests that more diverse arguments were communicated in Abortion hearings than in GMO hearings. This is consistent with our intuition because the policy of abortion is considered a more polarizing, 'hot topic' given the current political events surrounding this issue.

Republican majorities tend to invite more biased witnesses. Congressional scholars assume that hearings are a venue where committee members invite witnesses who would reflect the view of the members, especially the majority party members of the committee as they tend to control the selection of witnesses in most cases. To test this whether this belief holds empirically, utilizing our argument similarity tool, we calculate the similarity of arguments made by witness' in their testimony to members' arguments they made for a given hearing. This dyadic similarity was computed by members' party, whether their party held the majority status in the chamber, and by the policy issue area we consider. The difference in average argument similarity between majority/minority status was then used to test whether majority party members tend to make arguments similar to those conveyed in witness testimonies more than minority party members do. Evidence supporting this pattern would suggest that majority party members tend to invite witnesses who would reinforce their predispositions on a given policy issue. We use a t-test to examine this hypothesis.

Despite the absence of such pattern in hearings on GMOs, we found the statistically significant difference between when Republicans controlled the majority and thus largely managed the invitation of witnesses to discuss the abortion issue and when

clusters arguments.

Issue	Avg. Argumentativeness	# Arguments	# Pro	# Con
GMOs	0.1781	1548	208	80
Abortion	0.1265	3629	438	402

Table 1: Argumentation statistics for our data. The number of pros and cons are for or against the select topic with the most observations among the topics identified by WIBA-Extract in hearings on each policy issue.



Figure 5: The Level of Similarity Between Member and Witness Arguments in Hearings on Abortion

they were in the minority. The results are presented in Figure 5. This finding suggests that Republicans tend to be more strategic than Democrats in selecting witnesses who would testify in support of Republicans' own partisan views especially on highly partisan, salient issues, such as abortion, but these efforts become tenuous on issues that are less so, such as GMOs.

Abortion Q&A sessions reveal more new information than GMO hearings Q&A. In the literature on legislative studies, it is controversial whether any information acquisition actually happens during committee hearings. Theoretical works assume it does; empirical studies tend to see hearings are for public presentation of information the committees already obtained. We present a quick descriptive test on this controversy. We compute the pairwise cosine similarity scores for every argument made in the opening statements of a hearing to the arguments presented in the Question & Answer (Q&A) session. The distribution of scores is plotted in Figure 7 located in the Appendix, and we find the difference in similarity to be statistically significant. A higher similarity score between arguments indicates a duplicity of information presented, and therefore a lower similarity score is an indication of a *new* narrative being presented.

Furthermore, the same process is done with topics being argued in the opening statements, compared to the topics being argued in the Q&A session. These topics were identified using WIBA-Extract. A similar significant difference was observed, with more *unique* topics being argued in the Q&A session of Abortion hearings than GMO hearings, as shown in Figure 7 in the Appendix.

Abortion hearings are more polarizing than GMO hearings. Applying the WIBA-Extract and WIBA-Stance features, we compute the number of Pro and Con stance for a specific topic detected in each policy issue area we consider.⁴ For arguments specifically made about GMOs, we observe 208 Pro GMO arguments and 80 arguments against. For the Abortion topic we observe 438 Pro Abortion arguments and 402 arguments against.

We calculate the Controversiality Score for the arguments made in their respective hearings for these two topics to quantify the nature of discussion. For GMO hearings, we get a Controversiality Score of $\frac{80}{208} \times 100 = 38.5\%$. For Abortion hearings, we get a Controversiality Score of $\frac{402}{438} \times 100 = 91.8\%$. This result is intuitive given the polarizing partisan stances and American public on this policy issue.

⁴WIBA-Extract detects multiple topics conveyed in arguments made in each policy issue area. WIBA-Stance identifies whether the argument is for or against the given topic. Among multiple topics detected in Abortion hearing, we choose the topic with the most observations which was labeled as "abortion". Similarly, "GMOs" was the label of the topic with the most observations in GMO hearings.

This result also aligns with our findings of Deliberation Intensity, suggesting further investigation into the relationship between the two metrics.

Identifying Controversial Topics. Measuring the degree of controversy in discussion can provide useful insight into political discourse dynamics. We conduct our analysis by maximizing for the Controversiality Score, and additionally maximizing for the overall engagement, or total stances, for a topic. This helps to uncover topics that are controversial but also contain a significant amount of discourse.

First, we identify the most discussed topics overall within each of the two policy issues that we consider. For the GMO hearings, these are biotechnology, Genetically Modified Food, GMOs, organic farming, gene editing. For abortion, these are: Abortion, Planned Parenthood, right to life, Roe v. Wade, Hyde Amendment.

The most controversial and discussed topics for GMOs were Urbanization, GE Crops, dicamba, Roundup, and glyphosate, while the most controversial and discussed topics for Abortion were Planned Parenthood, right to choose, right to life, sex selection, Roe v. Wade, and Mifepristone.

Conversely, the *least* discussed topics that were controversial are the following for GMOs: Mandatory GMO labeling, bioengineering, GE Food, non-GMO food, and herbicide-resistant crops, and for Abortion: crisis pregnancy centers, ectopic pregnancy, Medicare, heartbeat legislation, judicial bypass procedures, and late-term abortion.

Most Argumentative Speakers Interestingly, for the Abortion hearings we find that Democrats make up the population with the highest argumentative score, whereas for GMO hearings witnesses make up the population with the highest argumentative score. This suggests that legislators tend to hold hearings to present their own arguments on hot button issues such as abortion whereas they tend to hold hearings to seek expert information and advice from field practitioners on technical issues such as GMOs. The following lists the five most argumentative speakers on each policy issue.

In Abortion Hearings: Jimmy Gomez (D), Richard Edmund Neal (D), a bureaucrat witness, Mark James DeSaulnier (D), Shontel Brown (D).

In GMO Hearings: Witnesses representing trade associations, corporate or non-profit organizations, Vicky Hartzler (R).

Interactive Dashboard to Navigate Arguments Made in a Hearing. To showcase our interactive dashboard that visualizes the arguments in the order they were presented by each speaker through a hearing session, we randomly selected one hearing on abortion, titled "Revoking Your Rights: The Ongoing Crisis in Abortion Care Access." This hearing was held by the House Committee on Judiciary on May 18, 2022, when the Democratic Party had the majority control over the chamber. Figure 3 presents how the interactive dashboard looks like. The horizontal axis records the number of sentences, which we labeled as "Time", and the vertical axis shows the last names of legislators with their partisan affiliation in parentheses and marks witnesses as "W" with their unique identifying numbers depending on the order they spoke for the first time in the hearing. The colored boxes represent clusters of unique arguments with the arguments in the same cluster or community, meaning the arguments that are similar with one another, is assigned the same color.⁵ Hovering over a bar in the chart reveals the argument and information on the person who said. Each text box below the chart summarizes the arguments made in the cluster to a single thematic argument. For example, the first three clusters from left to right are:

"The development of abortion bans has far-reaching consequences, exacerbating existing inequities and worsening health outcomes, particularly for women of color. By restricting access to abortion care, these bans can lead to increased maternal mortality rates, worsened economic outcomes, and generational harm."

"The development of a nationwide abortion ban is a clear attempt by Republicans to overturn state abortion laws and impose government control over women's bodies and choices."

"The fundamental right to control one's own body and make decisions about reproductive healthcare is a core American value, and no one deserves to be judged for seeking an abortion."

3.3 Conclusions

This paper showed how our WIBA framework can potentially be applied to analyze the domain of leg-

⁵The level of similarity used to determine the cluster or community can be manually adjusted up or down within the interactive dashboard.

islative deliberation in the U.S. Congress. We propose various useful metrics and visualizations that enables analyzing the nature of the discourse and the flow of arguments or reasoning on policy issues, which can be applied to future research on efficacy of legislative deliberation systems, legislators' behavior, influence of external groups on lawmaking processes, and even linguistic queries. We hope that these illustrations can stimulate thought on other measures and metrics of discourse that might be of interest to the fields of political science, public policy, sociology, psychology and linguistics.

References

- Arman Irani, Michalis Faloutsos, and Kevin Esterling. 2024a. Argusense: Argument-centric analysis of online discourse. Proceedings of the International AAAI Conference on Web and Social Media, 18:663–675.
- Arman Irani, Ju Yeon Park, Kevin Esterling, and Michalis Faloutsos. 2024b. Wiba: What is being argued? a comprehensive approach to argument mining. *Preprint*, arXiv:2405.00828.
- Thomas Jefferson. 1823. *Statutes at Large in Virginia* [1726], w.w. henin edition, pages 84–86.
- Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2021. Augmented sbert: Data augmentation method for improving bi-encoders for pairwise sentence scoring tasks. *Preprint*, arXiv:2010.08240.

A Appendix



Figure 6: The Level of Similarity Between Member and Witness Arguments in Hearings on GMOs.



Figure 7: (Left) The pairwise similarity distribution between arguments made in the opening statements to the arguments made in the Q&A session. (Right) Pairwise similarity distribution between topics argued about in opening statements to topics argued about in the Q&A session. These results indicate more repetition for GMO hearings, while there is an increase in knowledge for Abortion Hearings.