CSE 595 Project: Automatic Frame Classification in U.S. Congressional Speeches Using NLP Models

Olaf Dsouza
olafpd@umich.edu

Joshua Hsueh
jhsueh@umich.edu

Yana Patel
yrpatel@umich.edu

Abstract

Understanding how politicians frame policy issues is central to analyzing how leaders in the U.S. congress approach persuasion and agenda-setting on crucial matters in today's political landscape. However, manually classifying rhetorical frames in thousands of legislative text is labor-intensive and difficult to scale. In this project, we develop automated methods to classify rhetorical frames in Congressional speeches . We construct a dataset from the Congressional Record and assign multi-label frame annotations using few-shot LLM prompting with the Policy Frames Codebook. We evaluate two models: a TF–IDF LinearSVC classifier capturing lexical features, and a fine-tuned DistilBERT model capturing contextual semantics. Both models outperform a random baseline, with the SVM achieving the highest macro precision (0.6299) and DistilBERT achieving the highest macro recall (0.8498). These results indicate that Congressional framing exhibits strong stereotypical lexical patterns while also relying on deeper contextual cues. Our findings demonstrate that automated frame classification is feasible even under weak supervision. Future work includes long-context modeling, hierarchical architectures, and domain-adaptive contrastive pretraining to better capture the structure of extended political speech.

1 Introduction

The goal of this project is to develop a model that can automatically classify rhetorical frames in Congressional speeches. By doing so, we aim to enable large-scale analysis of how politicians across parties and over time strategically emphasize different dimensions of policy issues. Framing is the process by which speakers highlight certain aspects of an issue while downplaying other factors. For example, different frames for immigration include humanitarian focuses on morality and compassion, national security to protect against risks, and economic costs or benefits. Classifying the frames of a politician's stance is essential for the common person to understand the various political issues and their arguments. Furthermore, it can clarify to voters the different alignments and priorities of their politicians on national issues. This paper performs this analysis of political framing with recorded congressional speeches. One clear benefit from such a model is improving computational tools to study large-scale patterns of persuasion and agenda-setting by our Congress. Prior work shows frames can be measured reliably in news and political text. This paper hopes to expand this research into the Congressional record and compare it to simple, transparent baselines. In the end, our findings establish key foundations and notes for analyzing rhetoric in political persuasion and open up the door for automated analysis on the dynamics of Congressional speeches.

2 NLP Task Definition

The task this project tackles is multi-sentence level frame classification in Congressional speeches. The inputs would be text spans of multiple sentences, documented from the floor speeches in Congressional meetings. The output would consist of a set of frame labels. One good source of frame label classifications is the Policy Frames Codebook. Such frames include Economics; Capacity & Resources; Morality & Ethics; Fairness & Equality; Legality/Constitutionality; Crime & Punishment; Security & Defense; Health & Safety; Quality of Life; Cultural Identity; Public Sentiment; Political Factors; Policy Description/Prescription/Evaluation; External Regulation/International; Other. This will be a supervised learning task. Further analysis may include the trends of frames over time on big U.S. political issues.

3 Data

The primary corpus for the model is documented Congressional Record speeches readily available online. Our dataset is composed of speeches from the U.S. Congressional Record, which we accessed using the GovInfo API available at (https://gpo.congress.gov/). While the API provides structured metadata for each speech, such as the speaker, date, and chamber, we found that much of the full text was nested behind internal links to additional JSON endpoints. To handle this, we wrote a Python script that recursively followed these internal links, parsed the relevant text fields, and aggregated the speeches into a single corpus.

After gathering the data, we noticed that many of the entries contained a lot of noise. This included incomplete transcripts, vote records, and short procedural phrases like "I yield," "Without objection," or "Nay." To remedy this, we introduced a cleaning pipeline. It begins by removing entries that are less than 15 tokens, strips out any HTML tags or other unnecessary formatting pieces, and normalizes whitespace. The cleaning procedure also deduplicates records based on speaker name, date, and speech ID. In addition, we used a lightweight LLM filter that scans each speech and removes purely procedural sections or phrases that do not contain substantive policy discussion.

Following the cleaning procedure, our final dataset contains 13,999 total entries. Where 9,933 (71.0%) of them were usable speeches and 4,066 (29.0%) were ruled as procedural or repeated. Among the speeches, the average length is 2,407 tokens (median 1,208) per speech. There is a max length of 61,740 tokens and a min length of 22 tokens per speech. The dataset covers roughly 980 distinct speakers across both chambers, giving broad topical and speaker diversity.

Because no labeled frame dataset exists for the Congressional Record, we generated annotations automatically using an open-source GPT model with 20B parameters. We used a few-shot prompting setup based on the Policy Frames Codebook, which defines 15 high-level rhetorical frames. The model produced one or more frame labels per speech, and outputs were standardized to the 15-frame set. Figure 1 shows the distribution of the frames that were labelled by the LLM.

To verify the labelling quality, we also manually verified 15 randomly sampled annotated speeches. The LLM performed extremely well with our verifications, matching 14/15 of the annotated results from the model. For the one speech, although the frames were correct, we noticed the LLM took a broad approach to the labelling. Overall, our current dataset provides a foundation for modeling rhetorical framing in U.S. political speech.

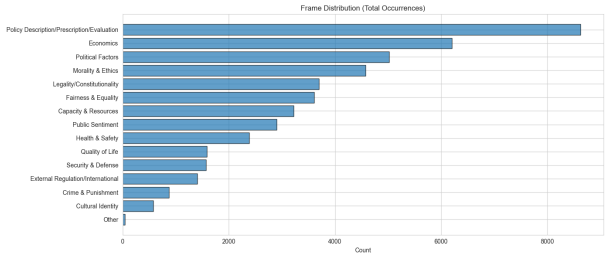

Figure 1: Distribution of Frame Labels in Dataset.

4 Related Work

4.1 Weakly Supervised Learning for Frame Tagging

Roy and Goldwasser (2020) wrote a research paper titled "Weakly Supervised Learning of Nuanced Frames for Analyzing Polarization in News Media". This research demonstrates that a weakly supervised embedding model is able to output more nuanced, topic-specific subframes in the news. They conclude that subframe indicators separate ideology better than coarse frame indicators. From this paper, we learned that it is possible for NLP models to learn classifications of frames in political contexts. This paper was limited to three U.S. issues while our project hopes to expand to more political issues and through politicians' Congressional speeches.

4.2 Comparison of Deep Learning Architectures to Classify Frames in News Headlines

Liu et al. (2019) investigate frames in news headlines about U.S. gun violence. This paper collected around 3,000 news headlines from various outlets and utilized a fine-tuned BERT (pre-trained Transformer) to classify these texts into nine frames. Not only did this model score well, but out-performed other baselines such as LSTM/GRU with attention using GloVe embeddings and Lexicon-based frame detection. The authors also were able to analyze the overall trends of frames over time, such as an increase in the mental health frame. This paper proves the effectiveness of using NLP models even in short texts to distinguish frames and identify

trends for key U.S. political issues.

4.3 Caution of Lack of Linguistic Theories Integrated in NLP Framing Tasks

Otmakhova et al. (2024) argues that many NLP research papers so far mistake topic classification with rhetorical framing. The authors create a typology of framing types and highlight gaps between linguistic/cognitive theories and NLP practice. Reviewing previous papers since 1997, they concluded equivalency, labelling, and narrative frames are underexplored. Furthermore, 70% of papers are too topic-stance focused with only 30% integrating valid linguistic theories. They emphasize a lack of integration of linguistic resources (e.g., FrameNet, narrative schemas). These distinctions are crucial as we work to classify frames with our own model.

4.4 Opinion Mining from YouTube Captions Using ChatGPT: A Case Study of Street Interviews Polling the 2023 Turkish Elections

Elmas and Gul (2023) confirms how LLM as a judge can be effectively used to classify frames from videos and speeches. The authors demonstrated how they could use GPT to produce initial frame labels (multi-label, not single class) as well as short evidence spans per frame. At the same time, the paper does caution against using the LLM labels as clear gold labels and encourages more as silver labels. For our project, we are validating the LLM labels with our own manual labels before automating.

4.5 UPPAM: A Unified Pre-training Architecture for Political Actor Modeling based on Language

Mou et al. (2023) demonstrates using pre-trained models such as BERT and fine-tuning them to capture embeddings for a political actor's speech style. This gave us some ideas for our potential methodology. It seems to be effective for capturing political styles to use pre-trained models. The paper also encourages us to consider using unsupervised or weak-supervised objectives on all unlabeled speeches to build better features before fine-tuning for frame classification. There may also be a benefit from using a representational learning stage where you train a sentence/paragraph encoder on large unlabeled speech corpora.

5 Methodology

To build our model, we began by converting each cleaned multi-sentence span into numerical feature representations suitable for supervised learning. For the baseline, we use a simple but robust bag-of-words approach with TF–IDF features combined with a Linear Support Vector Machine (SVM) classifier. We applied TfidfVectorizer from scikit-learn using unigrams and bigrams, a minimum document frequency of 5, and maximum document frequency of 0.9. We also enabled sublinear term frequency scaling to down-weight very frequent words and remove standard English stop words. Each speech span is therefore represented as a sparse high-dimensional TF–IDF vector. To capture contextual information beyond text content, we may also try to append additional metadata features. These are concatenated to the TF-IDF matrix using sparse vector stacking.

For classification, we trained a One-vs-Rest LinearSVC model, which fits one binary SVM classifier per frame label. Each SVM uses a regularization strength of C = 1.0 and class weights balanced inversely to label frequency. Because the standard SVM does not output probabilities, we calibrated its decision scores using a Platt-scaling sigmoid via CalibratedClassifierCV with three-fold cross-validation. This produces calibrated probabilities for each frame label, allowing us to fine-tune decision thresholds. We then sweep over a range of thresholds on the validation set to identify the optimal probability cutoff for each label that maximizes its individual F1-score. The final trained model outputs a binary vector of 15 frame predictions for every span as well as per-label probabilities. This model should capture strong lexical signals in Congressional rhetoric and excels in high-dimensional sparse text. Also due to its SVM architecture, this model should be a fast, interpretable, and competitive baseline for multi-label frame classification.

In addition to this baseline, we trained a neural model based on DistilBERT, a compact Transformer pretrained on English text, to capture deeper contextual semantics. We fine-tuned distilbert-base-uncased for multi-label classification by adding a single linear output layer mapping the 768-dimensional pooled embedding to 15 sigmoid-activated outputs (one per frame). The model was trained using binary cross-entropy loss (BCEWithLogitsLoss) with positive-class weights inversely proportional to label frequencies. We trained for

three epochs using the AdamW optimizer with a learning rate of 2e-5, a batch size of 16, and a weight decay of 0.01. Dropout of 0.1 and gradient clipping at 1.0 was also applied for stability. While also capturing strong signals in Congressional rhetoric, this model should be able to leverage contextualized Transformer embeddings to capture subtle rhetorical cues and frame patterns that cannot be detected using surface-level lexical features alone.

6 Evaluation and Results

To evaluate the model, the team measures performance on F1-score (both micro and macro). A validation and testing set was created separate from the training data for this evaluation. Some key baselines to compare to are random choice and existing LLMs such as GPT-5 and Gemini 2.5.

To set up the evaluation, the labeled dataset is divided into 70% for training, 15% for validation, and 15% for testing. Each baseline outputs a binary vector of predicted frames per span, along with the corresponding probabilities. Predictions were compared to the ground-truth LLM-as-Judge labels described earlier.

The first basic baseline is random choice. It randomly selects the number of frames to pick from a normal distribution centered at the average number of frames per speech (calculated from the dataset) and a standard deviation of 1.5. The frames are then chosen with equal probability out of randomness. This is then turned into a binary vector as the final prediction.

The next baseline is the LinearSVC model we created; the final baseline we want to compare against is based on DistilBERT.

Table 1: Random Baseline Evaluation Metrics on Congressional Speeches

Metric	Score
Number of Speeches Evaluated	9,933
Average Frames per Speech	4.66
F1 Score (Micro)	0.2923
F1 Score (Macro)	0.2538
F1 Score (Samples)	0.2680
Precision (Macro)	0.3094
Recall (Macro)	0.2769
Hamming Loss	0.4172

Table 2: Per-Frame F1 Scores for Random Baseline

Frame	F1 Score	Support
Policy Description/Prescription/Evaluation	0.4223	8631
Economics	0.3728	6207
Morality & Ethics	0.3515	4581
Political Factors	0.3508	5019
Fairness & Equality	0.3234	3608
Legality/Constitutionality	0.3144	3697
Capacity & Resources	0.3128	3223
Public Sentiment	0.2674	2903
Health & Safety	0.2619	2382
Security & Defense	0.2104	1574
Quality of Life	0.2006	1587
External Regulation/International	0.1767	1411
Crime & Punishment	0.1278	872
Cultural Identity	0.1065	578
Other	0.0077	43

Table 1 summarizes the overall random baseline performance across all Congressional speeches, serving as a simple point of comparison for later models. The baseline achieved a micro F1-score of 0.29 and a macro F1-score of 0.25, with an average of 4.66 frames per speech and a Hamming loss of 0.42. These results reflect the difficulty of the multi-label framing task under random prediction. Table 2 breaks down per-frame F1-scores, revealing that more frequently occurring frames such as Policy Description/Prescription/Evaluation and Economics achieved relatively higher scores, while rarer categories like Cultural Identity and Other had very low performance. This distribution highlights a label imbalance in the dataset.

Table 3 compares the F1 scores, precision, recall, and hamming loss across all three baselines (random, SVM, and DistilBERT). The random baseline performs poorly across all metrics, reflecting the inherent difficulty of predicting multiple rhetorical frames in long, heterogeneous political speeches. In contrast, both the TF–IDF and LinearSVC model achieves substantial gains, reaching micro-F1 scores of 0.7547 and 0.7424 as well as macro-F1 scores of 0.6592 and 0.6609 respectively. These improvements demonstrate that surface-level lexical cues can be highly predictive of framing choices in Congressional rhetoric. LinearSVC also produces the lowest Hamming loss of 0.1179, indicating strong overall accuracy in multi-label prediction.

DistilBERT, when fine-tuned end-to-end on our silver-labeled dataset, still achieves competitive performance. Although its precision is slightly lower than the linear model (0.5748 compared to

Metric	Random Baseline	SVM (TF–IDF)	DistilBERT (full)
F1 Micro	0.2923	0.7547	0.7424
F1 Macro	0.2538	0.6592	0.6609
F1 Samples	0.2680	0.5099	–
Precision (Macro)	0.3094	0.6299	0.5748
Recall (Macro)	0.2769	0.7026	0.8498
Hamming Loss	0.4172	0.1179	0.1303

Table 3: Model comparison across evaluation metrics.

0.6299), DistilBERT still attains the highest macro recall of 0.8498. This high result suggests that the neural model is more sensitive to subtle or infrequently expressed frames that may not exhibit strong lexical markers. This behavior is consistent with Transformers' ability to leverage contextual and semantic signals beyond simple word frequencies.

Taken together, these findings indicate that lexical features alone are sufficient to achieve strong framing performance, but contextualized neural representations provide complementary strengths, particularly in detecting minority or rhetorically nuanced frames. Both learned models substantially outperform the random baseline, confirming that rhetorical frame classification is a tractable and learnable task even under weak supervision.

Due to time constraints, we were unable to give a proper demo of our system working live to classify frames given any input speech for the poster presentation. A video demonstration of our model is found here: https://youtu.be/Awsh8RNWL-Y.

7 Discussion

As seen with the random baseline, classifying rhetorical frames in Congressional speeches is fundamentally challenging due to the highly multi-label and label-imbalance nature of political discourse. Several other issues exist such as individual speeches frequently invoking several overlapping frames. Another issue is that many of these frames are expressed implicitly rather than with explicit lexical markers. This complexity is exacerbated by label imbalance in the dataset, where a few dominant frames appear regularly, while others occur only rarely. Such imbalance makes high recall difficult for underrepresented frames and highlights the importance of models that can generalize beyond surface-level cues. Despite these obstacles, both the TF–IDF LinearSVC model and the DistilBERT classifier substantially outperform the random baseline, demonstrating that meaningful framing signals exist even in noisy, weakly supervised annotations.

A key observation from our evaluation is the complementary behavior of the two learned models. The SVM achieves strong macro precision, indicating that Congressional frames often correspond to relatively stereotyped lexical patterns that a linear classifier can capture effectively. DistilBERT, by contrast, displays the highest macro recall, suggesting that contextualized embeddings are more sensitive to subtle frame signals that span clauses or sentences. However, this improved recall also leads to reduced precision, implying that DistilBERT may overgeneralize on weak cues. The close performance between the two architectures highlights that although contextual models do provide additional capacity for nuance, Congressional framing is heavily driven by consistent lexical markers. This result makes classical linear methods more competitive than expected, suggesting that Congressional framing overall contains strong stereotypical lexical patterns. Taken together, this also suggests that the framing cues rely on both lexical patterns and deeper contextual semantics even cross-sentence.

Overall, these findings indicate that automated frame classification in legislative text is highly feasible using modern NLP techniques, even when trained on weakly supervised labels. The distinct strengths of linear and Transformer-based models suggest that an ideal framing system would integrate both lexical and contextual information to balance precision and recall. The results also provide evidence that political framing is structured and patterned enough for NLP model classification, opening up the door to scalable computational analyses of issue framing in the U.S. Congress. At the same time, the limitations observed in long-document handling, label ambiguity, and errors highlight important areas for methodological refinement.

8 Conclusion

This work demonstrates that automated classification of rhetorical frames in Congressional speeches is both feasible and informative, even when trained on LLM-as-a-judge. Examining both linear classifiers and contextual Transformer-based architectures, we find that Congressional framing follows consistent linguistic patterns that can be captured

effectively despite label imbalance, long-document complexity, and overlapping rhetorical strategies. While TF–IDF LinearSVC provides strong precision by leveraging stereotypical lexical cues, DistilBERT offers higher recall by capturing deeper contextual semantics. These results suggest that framing signals arise from both easy surface text as well as cross-sentence context and structure. These findings establish a foundation for scalable computational analysis of legislative rhetoric and open new opportunities for studying the dynamics of political persuasion in Congress. Future improvements in domain-adaptive representation learning and long context modeling promise to further enhance automated framing research and support richer political science.

9 Other Things We Tried

We began prototyping a hierarchical architecture where DistilBERT encodes sentences and a BiLSTM encodes the sequence of sentence embeddings. While theoretically this architecture can hold promise, the pipeline and augmentations needed for this to work proved significantly more complicated than anticipated. Due to the lack of time and effort, we were unable to train this kind of model with successful results in time for the project deadline. This idea was deferred to future work as described next.

10 Future Work (What We Would Have Done Next)

Based on the results and discussion of the current progress, there are a couple of future next steps to further improve the rhetorical analysis of Congressional records. A natural next step is to leverage our trained models to study the dynamics of rhetorical framing at scale. There is a lot of potential in longitudinal analyses that could reveal how frame usage shifts around major political events, policy debates, or elections. Similarly, comparative analyses across parties, chambers, or individual legislators could bring to light differences in rhetorical strategy. Such work has the potential to provide new insights into polarization, agenda-setting, and coalition formation within Congress.

A key challenge in the models explored so far in this paper is the extreme length of many Congressional speeches. These texts after normalization can often exceed the token limit of 512 in standard Transformer models. This forces aggressive truncation, leading to lost context that likely contributes to errors in frame detection. Possible remedies include long-context Transformers such as Longformer, BigBird, or RoPE-extended architectures, which can process several thousand tokens and preserve document-level coherence. Another promising direction is hierarchical modeling, where sentence-level encoders feed into document-level encoders. Such architectures align more naturally with the hierarchical structure of political speech and may yield in even better multi-sentence frame classification.

Given the size of the unlabeled Congressional Record, introducing contrastive pretraining on top of the DistilBERT architecture is another compelling avenue for improving model performance. In practice, we could train DistilBERT's political-speech encoders to differentiate between similar and dissimilar speech segments such as speeches across different issues, speakers, or ideological groups. Explicitly correlating these pairs, the model may learn embeddings that better capture the stylistic and rhetorical structure of legislative speech. These embeddings could then be fine-tuned for frame classification, potentially improving recall on rare frames and reducing reliance on the silver labels. Such improved architecture and analyses could significantly advance the state of automated rhetorical framing research.

11 Group Work

Olaf was mainly responsible for the data collection and API. Joshua was mainly responsible for the model architecture and design. Yana was responsible for gathering the results, figures, tables, and analysis of the results. Of course, we all helped each other out in the process. We all contributed to the related works and writing of the paper as well as poster presentation. We all equally contributed to the project.

References

Tuğrulcan Elmas and Ilker Gul. 2023. Opinion mining from youtube captions using chatgpt: A case study of street interviews polling the 2023 turkish elections. *ArXiv*, abs/2304.03434.

Siyi Liu, Lei Guo, Kate Mays, Margrit Betke, and Derry Tanti Wijaya. 2019. Detecting frames in news headlines and its application to analyzing news framing trends surrounding U.S. gun violence. In *Proceedings of the 23rd Conference on Computational*

Natural Language Learning (CoNLL), pages 504–514, Hong Kong, China. Association for Computational Linguistics.

Xinyi Mou, Zhongyu Wei, Qi Zhang, and Xuanjing Huang. 2023. Uppam: A unified pre-training architecture for political actor modeling based on language. In *Annual Meeting of the Association for Computational Linguistics*.

Yulia Otmakhova, Shima Khanehzar, and Lea Frermann. 2024. Media framing: A typology and survey of computational approaches across disciplines. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15407–15428, Bangkok, Thailand. Association for Computational Linguistics.

Shamik Roy and Dan Goldwasser. 2020. Weakly supervised learning of nuanced frames for analyzing polarization in news media. *CoRR*, abs/2009.09609.