Political Stance Classification of U.S. Congressional Bills

Imaad Khan

June 2025

Abstract

Understanding the ideological stance of legislation is crucial for analyzing political behavior and polarization. This project develops a model that classifies U.S. Congressional bills by both topic and political stance, using a multi-task Longformer-based architecture. Bill summaries were obtained from the U.S. Congress API and matched with roll call vote data from Voteview to derive NOMINATE dimension 1 scores, which act as a proxy for ideological stance. The model simultaneously performs topic classification (187 labels) and stance regression, trained on over 20,000 labeled examples. Results show high topic classification accuracy (80%) and reasonably strong stance predictions using NOMINATE-based labels. This project highlights the potential of NLP and deep learning to make complex legislative text computationally tractable, opening the door to automated political analysis, lawmaker ideology mapping, and policy trend monitoring.

1 Introduction

Legislation in the United States Congress is dense, nuanced, and ideologically driven—but often difficult for the public and researchers to interpret at scale. While voting records provide insight into lawmaker positions, few tools exist to directly analyze the political stance of the bills themselves. This project aims to fill that gap using natural language processing.

The model was designed to infer both:

- What the bill is about (topic classification)
- Where it sits politically (economic left vs. right stance)

This was not just a coding challenge, but a full pipeline of research, engineering, and experimentation—from raw bill titles to full summaries, data scraping, API limitations, label construction, and model training. The goal was to demonstrate how text models could be applied to understand congressional ideology and how lawmakers align on different policy topics.

https://github.com/ImaadKhan17/PoliticalModel

In an era of increasing political polarization and legislative complexity, tools that enable automated analysis of bill content and ideological leaning are more valuable than ever. This project not only enhances academic understanding of congressional behavior, but also provides a framework for using individual bills as proxies for lawmaker ideology. By analyzing how legislators sponsor and support specific texts, the model offers a scalable way to infer their stances on major issues—potentially making congressional behavior more transparent to the general public. These tools could help voters, journalists, and researchers better track how legislative actions align with public values and campaign promises.

2 Dataset and Labeling

Congressional Bills Project Dataset

The Congressional Bills Project dataset includes metadata and topic labels for over 20,000 U.S. Congressional bills, categorized into 187 distinct topics. However, it only contains titles and limited metadata, which lack sufficient detail for accurate text modeling.

U.S. Congress API Summaries

To enrich the input data, I used the U.S. Congress API to collect official bill summaries. A script was written to fetch summaries for each bill while respecting the API's rate limit of 5,000 requests per hour. These summaries provided essential linguistic detail to support model training.

Voteview NOMINATE Scores for Stance

To assign political stance labels, I used Voteview's NOMINATE dimension 1 scores, which range from -1 (liberal) to $+1$ (conservative). Each bill was matched with associated roll call votes, and an average of the sponsoring legislators' scores was used as a continuous stance label.

This process enabled multi-task learning with both discrete topic classification labels and continuous ideological stance values.

3 Model Architecture

The model architecture is based on the Longformer, a transformer optimized for long-document processing using sparse attention.

Multi-Task Setup

- **Topic Classification Head:** Outputs logits for 187 topic classes.
- **Stance Regression Head:** Outputs a continuous stance score between -1 and $+1$.

Input Processing

Each summary was tokenized and padded/truncated to a maximum of 1028 tokens. The Longformer produced contextualized token embeddings, from which a masked average was computed to obtain fixed-size representations.

Loss Functions

- **Topic Classification:** Cross-entropy loss.
- **Stance Regression:** Mean absolute error (MAE).

Both losses were weighted equally during training to optimize performance across tasks.

4 Training and Evaluation

The model was trained in Google Colab using PyTorch and Huggingface Transformers. The dataset was split into training and validation sets.

Optimization

The Adam optimizer with weight decay was used alongside a learning rate scheduler.

Metrics

- **Topic Classification Accuracy:** $\tilde{8}0\%$ on validation set.
- **Stance Regression MAE:** $\tilde{0}.15$.

Challenges

- API rate limits required careful batching.
- Missing data in Voteview introduced label noise.
- Long summaries required more memory-efficient models.

5 Results

Topic Classification

The model achieved 80% accuracy across 187 categories.

Stance Regression

The mean absolute error (MAE) was approximately 0.15, demonstrating reasonable performance despite noisy stance labels.

Example Predictions

- A healthcare bill received a liberal score, consistent with expectations.
- A defense bill received a conservative score, also in line with known partisan trends.

Insights and Limitations

- Using summaries instead of titles greatly improved accuracy.
- Multi-task learning improved both tasks.
- Some errors occurred on ambiguous or multi-topic bills.
- Conservative-leaning texts were harder to model accurately, likely due to dataset imbalance and linguistic differences.

6 Threshold Calibration for Stance Categorization

Since continuous stance scores are less interpretable, I binned predictions into:

- **Liberal:** score < -0.167
- **Centrist:** $-0.167 \leq$ score ≤ 0.025
- **Conservative:** score > 0.025

This threshold configuration, identified through a grid search of over 100,000 combinations, achieved 65.89% accuracy against a human-labeled set of 500 examples—up from 39% with a naive threshold of –0.1 to +0.1.

The slight left-shift in the centrist range suggests the model underpredicts conservative scores—possibly due to output asymmetry or training bias. This aligns with qualitative findings that the model interprets liberal language more reliably than conservative.

7 Conclusion and Future Work

This project presented a Longformer-based model for joint topic classification and political stance prediction of U.S. Congressional bills using bill summaries and roll call data. The results demonstrate the effectiveness of multi-task deep learning in modeling political text.

Future improvements may include:

- Using expert-labeled stance categories
- Improving conservative-language modeling with data augmentation
- Scaling to full bill texts
- Testing graph-based or encoder-decoder models
- Expanding to state and local legislation
- Building a public dashboard for real-time inference

Appendix

All code, data, and model files are available at: `https://github.com/ImaadKhan17/PoliticalModel`

References

- Voteview: Congressional Roll-Call Votes Database – `https://voteview.com/data`
- U.S. Congress API – `https://api.congress.gov`
- Beltagy, I., Peters, M. E., & Cohan, A. (2020). *Longformer: The Long-Document Transformer.* arXiv:2004.05150
- Liu, P., Qiu, X., & Huang, X. (2016). *Recurrent Neural Network for Text Classification with Multi-Task Learning.* arXiv:1605.05101
- Congressional Bills Project – Policy Agendas Project
- GitHub Repository: `https://github.com/ImaadKhan17/PoliticalModel`