Analyzing Nuclear Speech in U.S. Congress Records (2017-2024) using NLP Methods

Jihwan Lim*, Dong Hoon Lee, Keonhee Lee, Jiyoung Kim, Eunju Jun
Global Policy Research Section, Korea Atomic Energy Research Institute
**Corresponding author: limjh@kaeri.re.kr*

1. Introduction

Nuclear related issues are often presented as a rare area of bipartisan agreement in U.S. politics. Media narratives frequently emphasize cross-party consensus, pointing to examples like the ADVANCE Act of 2024, which passed both chambers of Congress with overwhelming support (House: 399-13; Senate: 88-2). While such legislative outcomes suggest unity, there has been little empirical investigation into whether this bipartisanship is reflected in actual congressional rhetoric. This is particularly true in the context of Natural Language Processing (NLP), where methodological tools now allow researchers to analyze large-scale political text data more efficiently than ever before.

Previously, political text data were often constrained by the massive volume of records, making them difficult to process without substantial resources. Due to high computational costs, only well-funded projects were able to make use of these resources [1]. Recent developments in NLP, especially the emergence of transformer based models, have made it feasible to conduct detailed analyses of text data such as legislative discourse at scale. These tools have already been applied to a range of topics related to the U.S. Congress, including polarization trends, agenda setting behavior, and rhetorical shifts, using both speech records and social media data [2, 3].

However, studies that specifically examine congressional discourse around nuclear issues remain outdated or limited in scope. Most prior work has focused on historical case studies or international contexts, such as U.S. relations with North Korea, and has relied primarily on qualitative methods. Despite the growth of NLP applications and renewed policy interest in nuclear energy, there remains a lack of systematic examination of how lawmakers discuss nuclear topics using computational text analysis.

This study seeks to fill the existing gap by using NLP methods to analyze Congressional Record data spanning the 115th to 118th Congresses. It focuses on identifying pro-nuclear and anti-nuclear rhetoric within the speeches of Democratic and Republican legislators, paying close attention to how these stances relate to perceived U.S. national interests. The selected period captures a complete range of partisan dynamics in Congress, including both unified and divided government scenarios across chambers. This diversity offers a solid basis for comparing institutional and party-based differences in tone. Table 1 outlines the partisan composition of Congress across these sessions.

Table 1: U.S. Congress Partisan Control

Congress	Years	House	Senate	President
115	2017-2018	Rep.	Rep.	Trump
116	2019-2020	Dem.	Rep.	
117	2021-2022	Dem.	Dem.	Biden
118	2023-2024	Rep.	Dem.	

Note: Because each Congress starts or ends on January 3 of an odd-numbered year, this study defines the congressional periods as ending on December 31 of the following even-numbered year for consistency in analysis.

2. Data Collection

This study uses full-text Congressional Record data, the official daily publication that records the activities and debates of the U.S. Congress, obtained from GovInfo, a public platform provided by the U.S. Government Publishing Office. All data covering the years 2017 to 2024 were collected using the **`congressional-record`** parser tool developed by the @unitedstates project on Github [4]. Consequently, the metadata was enriched by adding the year, party affiliation, and chamber using the *speaker_bioguide* field with reference to the **`legislators-current`** and **`legislators-historical`** that are also available on same project. The final base dataset includes 1,577,436 records, of which 421,673 are distinct individual speeches. The dataset was refined by removing procedural remarks and limiting entries to speeches by Democratic or Republican lawmakers, resulting in 274,099 speeches, with 49.4 percent from Democrats and 50.6 percent from Republicans.

To examine partisan differences before applying nuclear-specific filters, a keyword-based analysis was conducted on speeches by Democratic and Republican lawmakers, focusing on co-occurrences of the term nuclear with relevant policy keywords. Only speeches containing the word “nuclear” were included to ensure contextual relevance, and partisan asymmetries were measured using a log-ratio with add-one smoothing. The same method was applied to the full dataset as a baseline comparison, revealing that Republicans more

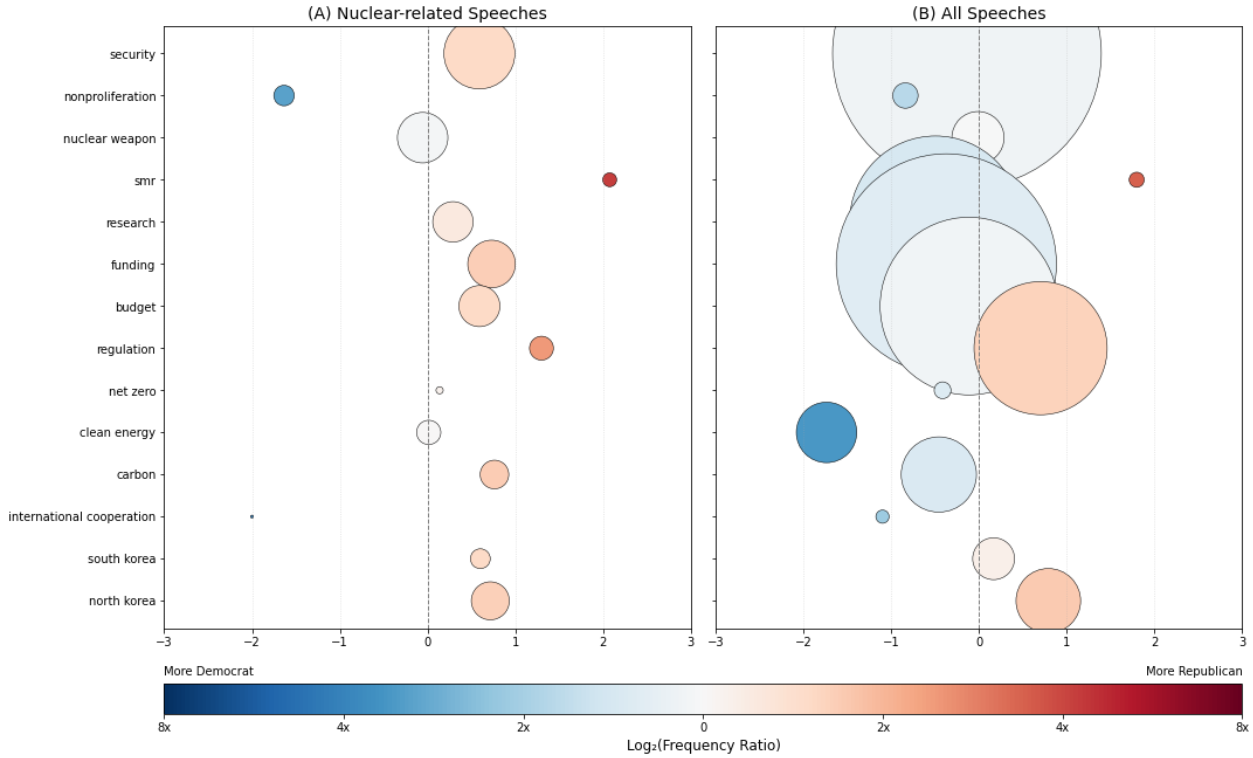

Fig. 1. Partisan usage of Terms. The left plot shows the relative usage of nuclear-related keywords in speeches that contain the word nuclear, while the right plot shows the same keywords across all speeches.

frequently referenced nuclear-related terms in nuclear-context speeches. Results are shown in Figure 1.

3. Preprocessing and Methodology

Nuclear-relevant discourse was identified through a multi-step preprocessing pipeline based on Aroyehun et al. (2025) [2]. The enriched dataset was filtered by stopword ratio threshold of 0.05 using the top 100 most frequent English stopwords from the Oxford English Corpus (OEC) and then segmented into 150-token chunks using a **`RoBERTa`** tokenizer, with short final chunks merged to ensure a minimum of 50-token chunks.

Keyword matching was applied to extract relevant content, using strict, loose, and context-sensitive terms. Only chunks meeting at least one criterion were retained, resulting in 4,102 chunks from 1,596 speeches by 413 lawmakers, balanced across party and chamber.

To continue, pro-nuclear, anti-nuclear, and neutral stances were extracted. Two stance classification methods were compared: a fine-tuned **`RoBERTa-base`** model and a zero-shot **`RoBERTa-large-MNLI`** model, both developed by Facebook AI (now Meta AI). Multiple hypothesis labels, ranging from simple to specific, were tested to improve zero-shot classification results. A total of 1,000 chunks were manually labeled, randomly and evenly sampled by party and chamber. The classification results, shown in Figure 2, indicate that the fine-tuned model performs more effectively.

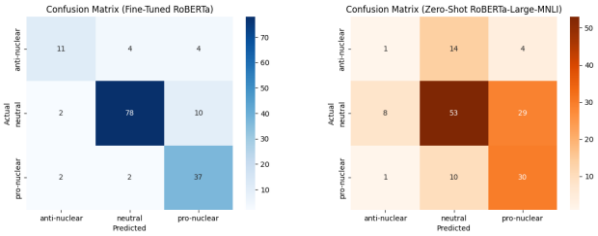

Fig. 2. Confusion Matrix of Fine-Tuned and Zero-Shot Models. The left plot shows the fine-tuned RoBERTa-base model, and the right plot shows the zero-shot RoBERTa-large-MNLI model. The dataset was split into training (70 percent), validation (15 percent), and test (15 percent).

4. Results

To evaluate partisan differences in tone over time, the net tone of nuclear-related speeches was calculated for each party and chamber from 2017 to 2024. Since the **`RoBERTa-base`** model outputs softmax probabilities for each stance category, each speech was assigned the label with the highest probability. Net tone was defined as the difference between the proportions of pro-nuclear and anti-nuclear speeches, scaled from –100 to 100. A score of 100 indicates all speeches were pro-nuclear, –100 indicates all were anti-nuclear, and 0 reflects an

even balance. Only speeches classified as pro-nuclear or anti-nuclear were included in the calculation.

As shown in Figures 3 and 4, a clear pattern emerges along both party and institutional lines. Figure 3 shows that Republicans consistently maintain a strongly pro-nuclear tone, with net tone values remaining above 50 throughout the period, though a decline is observed after the start of the Biden administration. In contrast, Democratic tone is more variable, dipping into negative territory around 2020 during the Trump presidency, then recovering in the early Biden years.

Figure 4 illustrates differences between the House and Senate. The Senate tone shows greater fluctuation, initially more pro-nuclear than the House, but this shifts after Democrats gain control in both chambers under Biden. The House tone remains more moderate and stable, with fewer year-to-year changes. A sharp peak in Senate tone in 2021 lacks a standard deviation band, likely due to a small sample size that year.

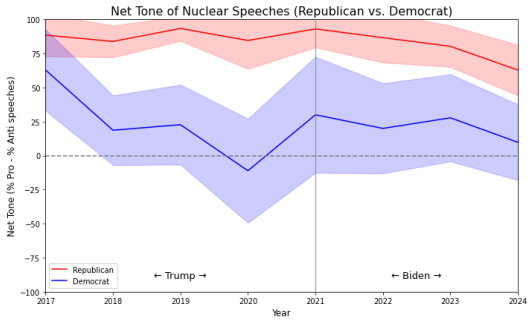

Fig. 3. Net Tone by Party. Red line represents Republicans and blue line represents Democrats. Shaded bands around each line represents ±2 standard deviations.

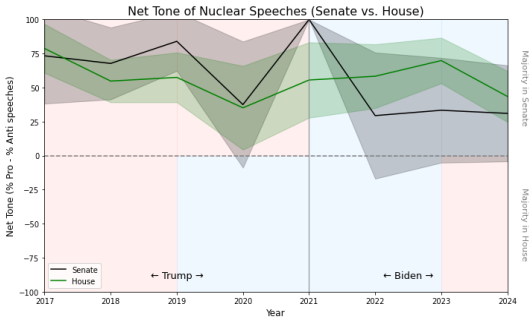

Fig. 4. Net Tone by Chamber. Black line represents Senate and green line represents House. Shaded bands around each line represents ±2 standard deviations. The background shading reflects partisan control of the period.

To formally assess the effects of party affiliation, chamber, and presidential administration on nuclear policy tone, a series of OLS regression models was conducted using speech-level net tone scores as the response variable. Net tone was calculated by linearly rescaling the pro-nuclear softmax probability into a continuous range between –100 and 100, using the formula $(2 \times p_{pro} - 1) \times 100$. For instance, a score of 0.0 corresponds to –100 (strongly anti-nuclear), 0.5 to 0 (neutral), and 1.0 to +100 (strongly pro-nuclear).

The predictors included party (coded as 0 for Democrats and 1 for Republicans), chamber (0 for House and 1 for Senate), and a binary variable for admin (presidential administration), with 0 representing the Trump years (2017–2020) and 1 the Biden years (2021–2024).

To account for heteroscedasticity, all Ordinary Least Squares (OLS) models were estimated using HC3 robust standard errors [5]. Model 1 examined the additive effects of party, chamber, and administration. Model 2 included an interaction between party and chamber to test for institutional moderation of partisan tone. Model 3 assessed whether the partisan tone gap varies across administrations, and Model 4 tested for shifts in chamber-level rhetoric across the same period. The continuous year variable was excluded to focus on administration-specific effects. Results are presented in Table 2.

Table 2: OLS Results

Predictor	**Model**			
	(1)	**(2)**	**(3)**	**(4)**
Intercept	-50.23*** (3.42)	-50.43*** (3.75)	-46.91*** (3.70)	-51.62*** (3.72)
party	37.57*** (3.89)	37.91*** (4.82)	31.48*** (5.04)	37.86*** (3.92)
chamber	-19.74*** (4.05)	-19.22*** (4.97)	-20.28*** (4.09)	-15.97** (5.16)
admin	11.20** (3.95)	11.24** (3.98)	4.03 (5.13)	14.06** (4.90)
party× chamber		-1.08 (8.22)		
party× admin			14.22 (7.93)	
chamber× admin				-8.98 (8.33)
Adj. R^2	0.074	0.073	0.075	0.074
F-statistic	47.80***	37.05***	35.96***	35.82***

Note: Robust standard errors (HC3) are in parenthesis. *p<0.05; **p<0.01; ***p<0.001.

5. Discussion

Congress plays a central role in shaping U.S. policy, including nuclear energy, weapons, and nonproliferation, with significant implications for national security and global leadership. While the legislative process has been extensively studied, the language used in congressional discourse on nuclear issues remains underexplored. This study offers a systematic attempt to analyze nuclear-related speech in Congress using NLP, applying a fine-tuned stance classifier to data from 2017 to 2024.

The results indicate that Republicans speak more often and more favorably about nuclear topics than Democrats, maintaining a consistently strong pro-nuclear tone across both energy and weapons. In contrast, the Senate shows a relatively less pro-nuclear stance than the House. Notably, pro-nuclear tone increased during the Biden administration among House Democrats, but this effect does not significantly vary across party or chamber. These findings suggest that even when bipartisan legislation like the ADVANCE Act of 2024 passes, rhetorical differences persist, reflecting deeper partisan divides in how nuclear policy is framed.

As a preliminary analysis, this study has several limitations. First, all 1,000 manually labeled samples were annotated by a single author, which raises concerns about subjectivity and reliability. This could be further improved by involving multiple annotators and reporting inter-annotator agreement using metrics such as Krippendorff's alpha. Second, keyword selection could have been strengthened through multiple rounds of validation by experts from relevant fields. Third, aggregating data at the yearly level flattened temporal variation and may have obscured short-term trends. Future work can improve these areas by incorporating collaborative annotation, expert-guided keyword refinement, and more granular time-based analysis.

REFERENCES

[1] J. Grimmer and B. M. Stewart, "Text as data: The promise and pitfalls of automatic content analysis methods for political texts," Political Analysis, vol. 21, no. 3, pp. 267–297, 2013.
[2] S. T. Aroyehun et al., "Computational analysis of US congressional speeches reveals a shift from evidence to intuition," Nature Human Behaviour, pp. 1–12, 2025.
[3] D. Card et al., "Computational analysis of 140 years of US political speeches reveals more positive but increasingly polarized framing of immigration," Proceedings of the National Academy of Sciences, vol. 119, no. 31, e2120510119, 2022.
[4] N. Judd, D. Drinkard, J. Carbaugh, and L. Young, congressional-record: A parser for the Congressional Record, Chicago, IL, 2017.
[5] J. S. Long and L. H. Ervin, "Using heteroscedasticity consistent standard errors in the linear regression model," The American Statistician, vol. 54, no. 3, pp. 217–224, 2000.