Predicting Convincingness in Political Speech: How Emotional Tone Shapes Persuasive Strength

Bhuvanesh Verma[1*] **Mounika Marreddy**[1*] **Alexander Mehler**[1]
[1]Goethe University, Frankfurt am Main, Germany
verma@em.uni-frankfurt.de, mmarredd@em.uni-frankfurt.de, mehler@em.uni-frankfurt.de

Abstract

Emotional tone plays a central role in persuasion, yet its impact on computational assessments of political argument quality in real world election campaign speeches remains understudied. In this work, we investigate whether positive emotional framing correlates with higher perceived convincingness in political arguments. We fine-tune language models on argument quality datasets and test their ability to transfer convincingness predictions to real-world campaign speeches. Using a corpus of U.S. presidential campaign speeches, we analyze emotional polarity in relation to predicted persuasive strength to test whether positively framed arguments are judged more convincing than neutral or negative ones. Our empirical analysis shows that political parties rely heavily on argumentation during their election campaigns. Also, we found the evidence that politicians strategically employ emotional cues within their arguments during these campaign speeches, with positive emotions being more strongly associated with persuasive strength, for example in topics such as *USMCA's Effect on American Jobs and Agriculture*, *Border Control Policies*, *Progressive Tax Reforms*. At the same time, we find that negative emotions have a weaker yet still non-negligible influence on voter persuasion in topics such as *City Crime and Civil Unrest* and *White Supremacist Violence (Charlottesville Incident)*.

1 Introduction

Persuasion modeling estimates how convincing language is based on its linguistic and rhetorical features. Prior political psychology research shows that emotional framing significantly shapes information processing, attitude formation, and belief revision (Petty and Cacioppo, 2012; Early, 2015; Lerner et al., 2015; Rodriguez-Hernandez et al., 2024). Positive affect is linked to greater message acceptance, trust, and credibility in persuasion (Shah, 2022; Hassell et al., 2022; Grüning and Schubert, 2022). Despite the known link between emotion and persuasion, it remains unclear whether emotional tone consistently affects the perceived convincingness of political arguments and whether modern Natural Language Processing (NLP) models capture this in large-scale political discourse. To address this gap, we hypothesize that positively framed arguments are predicted to be more convincing than neutral or negative ones.

Computational persuasion research has largely focused on modeling argument quality, stance, and rhetoric across debates, online discussions, and social media (Marreddy et al., 2025; Lippi and Torroni, 2016; Haddadan et al., 2019; Tan et al., 2016; Qiu et al., 2015). In contrast, campaign speeches, which are key to electoral messaging and identity building, remain understudied computationally. Unlike debates, which are adversarial and episodic, campaign speeches are narrative and affective, aimed at sustaining momentum and alignment across elections (Ross, 2006; Chong and Druckman, 2007). Large-scale analyses reveal systematic variation in emotional rhetoric across political actors and time (Gennaro and Ash, 2022; Aroyehun et al., 2025), yet most focus on sentiment, lexical patterns, or themes rather than persuasion outcomes (Finity et al., 2021; Mohapatra and Mohapatra, 2022). Thus, the link between emotional tone and persuasive strength in campaign speech remains underexplored.

Advances in argument mining and persuasion modeling make it possible to study such questions at scale (Lawrence and Reed, 2019; Wachsmuth et al., 2016; Druckman, 2022). Neural language models in particular have demonstrated strong performance in predicting argument convincingness and agreement across domains (Lawrence and Reed, 2019; Wachsmuth et al., 2016; Druckman, 2022). However, whether these models exhibit

[*]Equal contribution.

The Proceedings for the 15th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis (WASSA 2026), pages 37–51
March 29, 2026

stable emotion-sensitive persuasion patterns in political speech remains unknown.

To investigate this question, we examine persuasive framing in 2020 U.S. presidential campaign speeches using Chalkiadakis et al. (2025)'s corpus of 1 056 Democratic and Republican speeches (Jan 2019–Jan 2021). We extract topic-aligned arguments and assess their predicted convincingness using models fine-tuned on persuasion and argument-quality datasets. We also examine how emotional polarity, especially positivity, relates to predicted persuasiveness across candidate, topic, and model scale.

We make three contributions: (1) We provide the first large-scale empirical analysis examining how emotional tone interacts with predicted persuasive strength in real-world campaign speech. (2) We evaluate whether persuasion modeling trained on benchmark datasets transfers to political speech through significance testing on downstream analysis. (3) We present a comprehensive empirical analysis showing that both parties rely heavily on argumentation in campaign speeches, with no significant differences in overall argument usage or aggregate argument quality, but with clear topic-specific asymmetries in persuasive effectiveness driven by emotional cues. In particular, we show that positive emotions are more strongly associated with persuasive strength, while negative emotions also play a non-negligible role in persuasion for certain topics.

2 Related Work

NLP has explored persuasion in political communication through argument mining, emotional rhetoric, and convincingness prediction. Early argument mining emphasized structure over persuasive impact in political communication. For example, Lippi and Torroni (2016) found that prosodic and lexical features improve claim detection in debates, while Menini et al. (2018) showed that argument mining reveals ideological contrasts in presidential speeches. Large-scale resources like the 29K-component corpus from Haddadan et al. (2019) enabled systematic modeling of argument structure across 50+ years of U.S. presidential campaigns.

Related literature examines emotion in political speech. It shows systematic variation in sentiment and affect across parties, candidates, audiences, and campaign contexts (Mohapatra and Mohapatra, 2022; Gennaro and Ash, 2022; Aroyehun et al., 2025). However, these studies mostly focus on descriptive emotional patterns rather than how emotional tone shapes persuasive outcomes. Communication theory and psychology suggest that emotional framing affects message acceptance, credibility judgments, and belief updating (Petty and Cacioppo, 2012; Lerner et al., 2015; Shah, 2022). Similarly, computational studies link emotion to stance and engagement (Benlamine et al., 2015, 2017), with recent work revealing modeling biases, such as overpredicting negative emotion in argumentation (Greschner and Klinger, 2024).

A third line of work focuses on computational persuasion modeling and convincingness prediction. Early work introduced pairwise ranking frameworks to estimate persuasive strength (Habernal and Gurevych, 2016), later improved through neural architectures capable of capturing stylistic and discourse level signals (Chowanda et al., 2017). Cano-Basave and He (2016) investigated the argumentation styles of speakers in political debates and modeled persuasion using audience reactions such as applause and booing as proxies for emotional response. Similarly, Quensel et al. (2025) investigated the influence of subjective factors, including emotions, on argument strength, with a particular focus on persuasion.

Persuasion has also been modeled in online debates and stance settings, including forecasting agreement shifts and outcome prediction (Tan et al., 2016; Qiu et al., 2015). Yet little work tests whether persuasion-trained models capture emotional framing, especially in campaign speech, or whether effects generalize across topics, speakers, or model scale.

Unlike this prior work, we examine how emotional tone relates to predicted persuasiveness in real-world U.S. campaign speeches and whether persuasion-trained models reflect emotion-sensitive persuasion patterns. In addition, we evaluate these effects across multiple model families and speaker contexts, allowing us to test the robustness and generalizability of emotion–persuasion interactions beyond the domains typically considered in earlier computational persuasion studies.

3 Datasets Description

We use three datasets in this study: (1) an argument detection corpus used to identify argumentative text spans, (2) an argument quality dataset used to train convincingness models, and (3) a campaign speech

corpus used for downstream analysis of persuasive framing in political communication.

3.1 Argument Detection Dataset

In this work, we adopt a broad definition of argumentation. A statement is considered argumentative if it expresses reasoning, justification, or support within a text, regardless of whether it functions as a claim or as evidence. Accordingly, both claims and evidence are treated as instances of argumentative language. To operationalize this definition, we combine two existing argument detection datasets, Ein-Dor et al. (2020) and Levy et al. (2018), to construct the ArgDetect dataset. One dataset provides claim–topic pairs, while the other consists of evidence–topic pairs; we merge these resources to create a unified argument detection corpus. Dataset statistics are reported in Appendix A and Table 2.

3.2 Argument Quality Dataset

For training a model to detect argument quality, we used EviConv by Gleize et al. (2019) which is a relative evidence ranking dataset. This dataset contains a topic claim and two evidence statements and a label. Label assigns which evidence is better among the two. Since our objective is to conduct a comparative analysis of argumentative strength between political parties, this relative formulation is a natural fit for our use case. In addition, adopting a ranking-based framework reduces sensitivity to dataset and modeling calibration issues that commonly arise when transferring absolute convincingness scores across domains (Ivanova et al., 2024). The dataset statistics are presented in the Appendix A Table 3.

3.3 Campaign Speeches Text Data (CSTD)

For our experiments, we use the Campaign Speeches Text Data (CSTD) introduced by Chalkiadakis et al. (2025) and released in Scientific Data. The corpus comprises 1,056 campaign speeches delivered by the Democratic and Republican presidential and vice-presidential candidates during the 2020 U.S. election cycle, spanning January 2019 to January 2021.

4 Methodology

Rhetorical analysis of political speeches involves identifying rhetorical components, such as arguments, and quantifying them. To this end, we first identify the latent topics present in campaign speeches, followed by the extraction of argumentative statements and the evaluation of their persuasive quality and associated emotional strength. Therefore, we train models for argument detection and argument quality assessment and along with that topic extraction is performed using unsupervised topic modeling technique, while emotion classification is conducted using *SamLowe/roberta-base-go_emotions* [1] which was trained on the GoEmotions dataset that includes 27 distinct emotion labels. In the following subsections, we describe each methodology in detail.

4.1 Argument Detection

For argument detection, we train traditional machine learning (ML) models such as *Logistic Regression*, *Linear SVM*, and *Naive Bayes*, transformer-based models including *bert-base-uncased*, *roberta-base*, and *deberta-v3-large*, and large language models (LLMs) like *Qwen2.5-Coder-14B*, *Mistral-8B-Instruct-2410*, and *Llama-3.1-8B-Instruct*. The dataset is split into training and test sets in an 80:20 ratio for all experiments. For ML and transformer-based models, we concatenate the topic title with the sentence, separated by a space, to form each training instance. With ML-based models, we use TF-IDF and sentence embeddings (Reimers and Gurevych, 2019) (SE) features based on unigrams and bigrams (maximum 10,000 features) with stopword removal. For transformer-based modeling, we added a classification head on top of each pretrained model for sequence classification. To address the natural class imbalance of argumentative versus non-argumentative sentences, we apply a class-weighted cross-entropy loss (Phan and Yamamoto, 2020). Model training uses a learning rate of 2e-5 and a weight decay of 0.01.

For efficient utilization of LLMs, we evaluate their two-shot prompting capabilities, as zero-shot settings often struggled to produce correctly formatted labels. The prompt used for two-shot evaluation is provided in the Appendix B Section B.1. To optimize inference efficiency, we employ constrained generation (Beurer-Kellner et al., 2023), restricting the LLMs to output a single token corresponding to the label (0 or 1). During generation, we consider the probabilities of these two tokens to determine the predicted label. For evaluation, we randomly sample 200 instances from the test set, calculate

[1] https://huggingface.co/SamLowe/roberta-base-go_emotions

metrics, and repeat the experiment five times to ensure stability.

4.1.1 Argument Quality Assessment

Similar to the argument detection task, we train a set of traditional machine learning (ML) models, transformer-based models, and evaluate two-shot capabilities of LLMs. We divide the training set of EviConv to create a validation split using a 9:1 ratio. For ML models, the topic and its two associated evidence sentences are concatenated into a single input text. Similar features and preprocessing steps are applied as those used in the argument detection modeling. For transformer-based modeling, we implement a Topic-Evidence Ranker that scores evidence sentences relative to a given topic. Each claim-evidence pair is encoded using a shared pretrained transformer (e.g., *bert-base-uncased*), and the resulting [CLS] token embedding is passed through a two-layer MLP with a hidden size of 256 to produce a scalar relevance score. During inference, evidence sentences are ranked according to these scores, with higher scores indicating more persuasive argument. The model is trained using pairwise ranking loss for three epochs to distinguish the more convincing evidence. For LLM evaluation, we adopt the same two-shot prompting approach, providing two-shot examples in the prompt (see prompt in the Appendix B Section B.2).

4.1.2 CSTD Analysis

With the argument detection and argument quality ranking models in place, we first extract topics from the CSTD using *BERTopic* (Grootendorst, 2022). To prepare the data, we aggregate clean text from all speeches across all speakers into a single corpus. The text is then chunked into segments of 512 tokens to prevent truncation, and each chunk is embedded using sentence embeddings (*all-MiniLM-L6-v2*) and passed to BERTopic. This setting allow us to capture the whole argument within a text chunk as an argument can span to multiple sentences. For the generated topics, we assign labels using LLMs. Specifically, multiple LLMs including *Gemma3* (Team et al., 2025) and *Qwen2.5* (Qwen et al., 2024) are prompted to suggest labels, and a third LLM (*Llama3.2*) is used to align and finalize the topic labels. Once topics are established, we obtain the corresponding sentences for each topic, which serve as input for the argument detection model. Each topic-sentence pair is processed to identify argumentative statements within the CSTD. Subsequently, arguments are grouped by political affiliation Republican or Democratic. To compare arguments between parties, we form pairs of opposing arguments for each topic. These pairs, along with the associated topics, are then evaluated using the argument quality ranking model to determine which argument is more persuasive. This process yields a ranked list of the most compelling arguments for each topic, along with their associated political party.

Finally, to examine the emotional content of the identified arguments, all argumentative statements are passed through an emotion classification model. This enables analysis of the nuanced emotional tone present in persuasive political statements.

5 Results and Discussion

In this section, we present, results from training argument detection and argument quality model. We also present the analysis on Campaign Speech dataset including topic modeling and argument prediction. Finally, we analyse the results with respect to political affiliation, argument convincingness and emotions involved.

5.1 Argument Modeling Results

Transformer-based models achieved the strongest performance across both the argument detection and argument quality assessment tasks. In argument detection, our primary objective was to maximize precision in order to identify true argumentative statements with higher reliability. Among ML models, Linear SVM with pretrained sentence embeddings performed notably well, achieving a precision of 0.77 and an F1-score of 0.74. In the two-shot LLM setting, *Llama-3.1-8B-Instruct* demonstrated competitive performance, suggesting that parameter-efficient fine-tuning (e.g., PEFT) could further enhance results. However, transformer-based models particularly *DeBERTa-v3-large* achieved the best performance, with a precision of 0.87 and an F1-score of 0.86 (see full results in Table 1). This model was therefore selected for extracting arguments from CSTD.

For argument quality assessment, we prioritized accuracy, as the goal was to reliably determine the more persuasive argument in each pair. ML models again benefited substantially from sentence-embedding features, which outperformed traditional TF-IDF representations. Among LLMs, *Ministral-8B-Instruct-2410* achieved the highest ac-

Model	ArgDetect		EviConv	
	Precision	F1	Accuracy	F1
Logistic Regression (tf-idf)	0.78	0.72	0.50	0.48
Linear SVM	0.75	0.74	0.47	0.46
Naive Bayes	*0.79*	0.59	0.52	0.48
LR with SE	0.77	*0.74*	*0.54*	*0.52*
BERT-base-uncased	0.84	0.84	0.75	0.76
RoBERTa-base	0.84	0.84	**0.78**	**0.79**
DeBERTa-v3-large	**0.87**	**0.86**	0.78	0.78
Qwen2.5-Coder-14B	0.6321 ± 0.0211	0.4581 ± 0.0169	0.616 ± 0.0305	0.6166 ± 0.0305
Ministral-8B-Instruct-2410	$\mathit{0.6473 \pm 0.0077}$	0.5098 ± 0.0154	$\mathit{0.658 \pm 0.0223}$	$\mathit{0.6573 \pm 0.0223}$
Llama-3.1-8B-Instruct	0.6353 ± 0.0196	$\mathit{0.6162 \pm 0.0219}$	0.64 ± 0.045	0.6350 ± 0.0466

Table 1: Results of ML, Transformer, and LLM models on ArgDetect and EviConv, with macro-averaged metrics. For LLMs, mean and standard deviation are from five two-shot samples of 200 instances. Italics mark the best ML model; bold marks the best overall per dataset.

curacy (0.65). Transformer-based models such as *RoBERTa-base* and *DeBERTa-v3-large* performed comparably, and our Topic–Evidence Ranker improved upon the baseline accuracy of 0.73 (Gleize et al., 2019) across all pretrained model. For downstream analysis of CSTD, we selected the *RoBERTa-base* ranker (accuracy 0.78) for argument quality assessment.

5.2 Topic Modeling Results

After applying topic modeling to the CSTD, we identified 88 interpretable topics (see full list in the Appendix C). Although several topics exhibited semantic proximity, we retained them as distinct categories to preserve thematic granularity in downstream analysis. Across these topics, we extracted 11,944 topic-relevant sentences, of which 73% were produced by Republican candidates and 27% by Democratic candidates. Notably, 72 of the 88 topics were discussed by speakers from both parties.

The most frequently discussed topics across all campaign speeches were *Border Control Policies*, *Small Business Policy, and Political Alliances* and the *USMCA's Impact on American Jobs and Agriculture*. Party-specific rankings reveal both convergence and divergence. For Republicans, the top topics include *Border Control Policies* and *USMCA's Effect on American Jobs and Agriculture*, both of which appear in the overall top three, as well as *U.S. Election: Allegations of Government Corruption and Family Scandals*. In contrast, the Democratic Party's top three topics are markedly different: *Universal Education Equity*, *Universal Access to Comprehensive Healthcare (Including Pre-existing Conditions)*, and *American Values Advocacy: Constitution, Law Enforcement, Faith, Family, and History*.

Overall, Democratic speakers tend to emphasize the expansion of social programs and civil rights, including universal healthcare and education, gun control, voting rights, and climate policy through clean-energy initiatives. They also highlight pandemic-related economic relief for individuals and small businesses. Republican speakers, by contrast, prioritize national security, traditional values, and limiting federal intervention. Their discourse frequently centers on border control, "American values" (e.g., constitutionalism, faith, and family), conservative judicial appointments (such as pro-life positions regarding the Supreme Court), investigations into alleged government corruption, and election integrity. Additionally, Republicans devote attention to trade-related economic effects (e.g., USMCA) and veterans' affairs.

5.3 Political Argument Detection

Using trained argument detection model (see subsection 4.1), we extract arguments used by the politicians in the campaign speech data irrespective of the fact if its a claim or evidence. Overall 36.84% of the total text chunks were arguments. Republican vice presidential candidate Mike Pence emerged as one with highest argument percentage while Donald Trump with least argument percentage as 34.81%.

Within the Democratic party, Joe Biden produced a higher proportion of arguments than Kamala Harris. Certain topics demonstrated particularly high argument rates, including *Tax Relief for Family Farms and Small Businesses*, *COVID-19 Testing Capacity Growth (US, Feb '20)*, and *Job Dignity & Workplace Respect*, which ranked

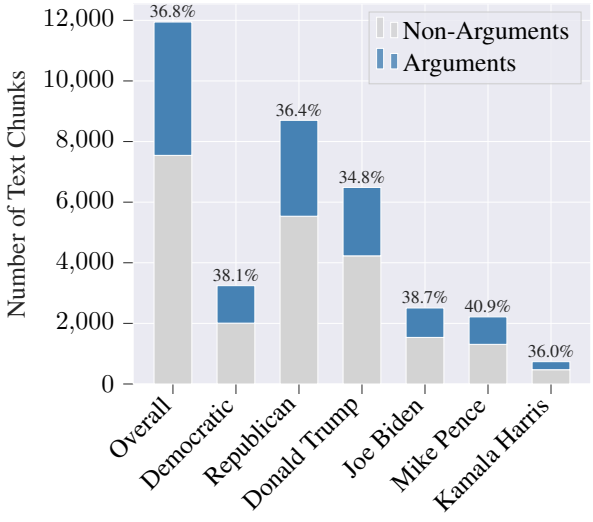

Figure 1: Distribution of predicted arguments in the campaign speech dataset. The stacked bars show the number of argument (steelblue) and non-argument (light gray) text chunks for the overall dataset, broken down by party, and by individual candidate.

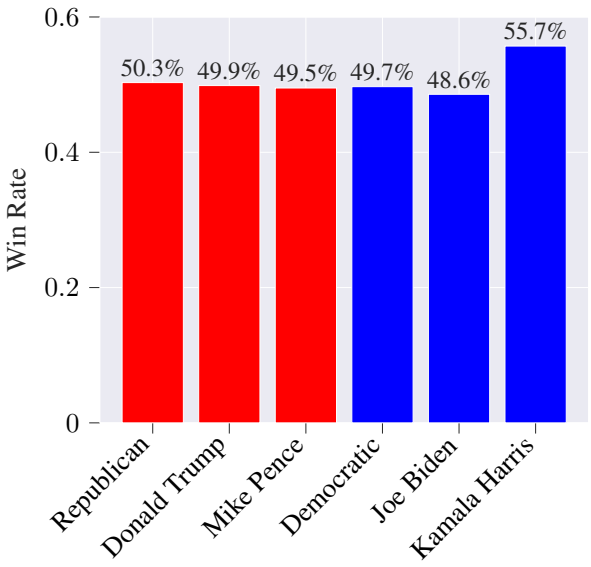

Figure 2: Pairwise convincingness by party and speaker. Bars show win rates, with colors indicating party affiliation (red = Republican, blue = Democratic).

as the top three in argument density. Conversely, topics such as *U.S. Election: Allegations of Government Corruption and Family Scandals* and *America Reclamation Movement* exhibited lower argument rates. As illustrated in Figure 1, Democratic speeches contained a higher percentage of arguments overall compared to Republican speeches. It should be noted, however, that the total number of text chunks was greater for Republican speakers, resulting in a higher absolute number of arguments despite the lower proportion.

5.4 Political Argument Quality Analysis

For analyzing argument quality, we automatically paired arguments for each topic between politicians from different parties. For each topic, we extract arguments are extracted independently for each party, after which all possible cross-party argument pairs are generated exhaustively. This setup allows a direct comparison of argumentative strength between parties and aligns naturally with our argument quality model, which is designed to perform relative comparisons between arguments. As shown in Figure 2, the overall distribution of convincing argument pairs is approximately balanced between the two parties. Examining individual performance, Kamala Harris achieves the highest overall convincingness rate, outperforming both Donald Trump and Mike Pence. When examining convincingness at the topic level, Democratic arguments exhibit a substantially higher overall rate: 65.5% of their arguments are judged more convincing, compared to 34.4% for Republican arguments. This suggests that, despite variation across individual topics, Democratic speakers generally produce more persuasive arguments. An overall semantic map illustrating topic-level argument convincingness is presented in Figure 3.

5.5 Emotions in Persuasive Political Discourse

In the previous section, we presented pairwise results of argument convincingness. To examine the role of emotions in persuasive political discourse, it is necessary to convert these pairwise results into a single, argument-level convincingness score. To achieve this, we calculate the *argument convincingness rate* by counting the number of times a given argument is selected as the winning argument. These arguments are then processed through an emotion detection model to obtain both the emotion category and corresponding intensity.

Our analysis shows that more than half of the arguments (56.5%) are classified as **neutral**. Beyond neutrality, positive emotions such as admiration, approval, gratitude, and optimism are most frequently associated with arguments. Among negative emotions, disapproval occurs most frequently, followed by sadness, disappointment, and annoyance. Interestingly, although only two arguments are associated with anger, they exhibit high convincingness rate (97.3%). Other negative emotions,

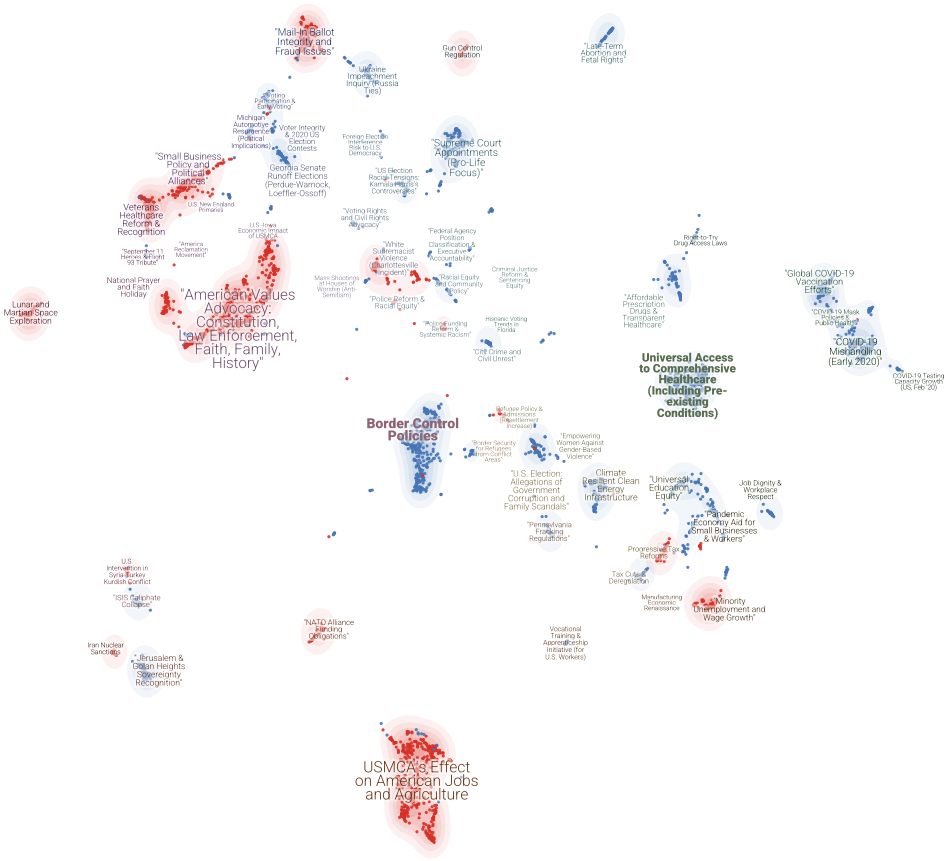

Figure 3: Topic-Convincingness Map for CSTD with blue representing Democratic and red representing Republican

including fear, disapproval, and annoyance, also correspond to arguments with high convincingness. For positive emotions, curiosity, love, amusement, desire, and surprise are among those linked to particularly persuasive arguments. Figure 4 shows distribution of the most dominant emotions across topics in CSTD.

In Republican arguments, the topics with the highest convincing rates were *City Crime and Civil Unrest* (100%) and *Pennsylvania Fracking Regulations* (92%). For *City Crime and Civil Unrest*, emotions observed included admiration, sadness, love, annoyance, and disappointment, with intensity analysis showing that **sadness and annoyance** were more pronounced than admiration and love. For *Pennsylvania Fracking Regulations*, disapproval and admiration appeared with comparable intensity. Examining the most-discussed Republican topics, such as *USMCA's Effect on American Jobs and Agriculture* and *Border Control Policies*, we found that **admiration, approval, and gratitude** were prevalent, with admiration and gratitude exhibiting the highest intensity.

In Democratic arguments, the most persuasive topics were *White Supremacist Violence (Charlottesville Incident)* (98%) and *Progressive Tax Reforms* (77%). For the Charlottesville Incident related topic, key emotions included fear, approval, and sadness, with **fear and sadness showing high intensity**, though neutral statements constituted 83% of the total. For *Progressive Tax Reforms*, the dominant emotions were approval and curiosity. Looking at the most-discussed Democratic topics, such as *American Values Advocacy*, *Universal Access to Comprehensive Healthcare*, and *COVID-19 Mishandling (Early 2020)*, emotions varied by topic. In *COVID-19 Mishandling*, sadness and gratitude were prominent, with gratitude exhibiting high intensity. In healthcare-related discussions, approval predominated, accompanied by other emotions such as sadness and curiosity.

Overall, the analysis indicates that politicians employ emotions strategically in their argumentation. For example, Democrats leverage negative emotions such as fear and sadness when discussing topics related to white supremacist violence, aligning their messaging with public concern on these issues. In contrast, when addressing *Progressive Tax*

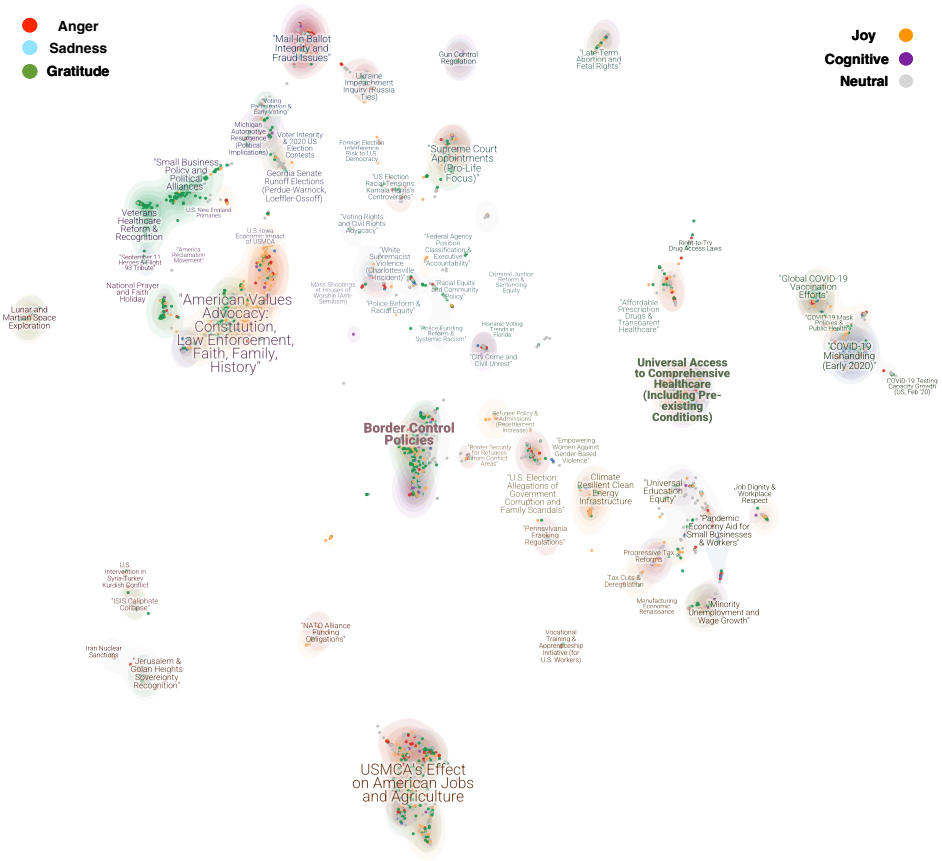

Figure 4: Topic-Emotion Map for CSTD. Emotional tone used by both political parties across various topics identified in the CSTD.

Reform, they predominantly use positive emotions like approval to appeal to voters who support taxing the wealthy. Similarly, Republican candidates utilize positive emotions such as admiration and approval when discussing the USMCA trade agreement, whereas for topics like city crime and civil unrest, particularly in states with active protests, they employ negative emotions to resonate with their voter base.

5.6 Randomization Based Significance Tests

To assess the statistical significance of our findings, we conducted a series of randomization tests. First, we examined argument detection results. By maintaining the original proportion of text excerpts per party, we randomly reassigned excerpts to each party. A direct comparison between Democrats and Republicans showed no significant difference relative to the random baseline, indicating that the observed difference in argumentative text proportions could occur by chance.

Next, we investigated argument quality, measured via convincingness scores. All arguments were randomly reassigned across parties while keeping the number of arguments per party constant, and average convincingness was computed in each randomization. The observed results indicated comparable convincingness between parties, with Republicans exhibiting a slight advantage (0.5%). Randomization testing, however, revealed no evidence of a significant overall pairwise advantage for either party.

We assessed topic-level argument comparisons. For each topic, multiple pairwise comparisons between arguments were available, with the winner of each comparison recorded. Randomization was applied within topics by shuffling winners across comparisons, and the resulting distributions were compared to the observed win ratios. This analysis identified 10 topics in which Democratic arguments exhibited a significant edge, indicating a topic-specific advantage in convincingness. Similarly, 15 topics were found in which Republican arguments showed significant superiority.

Finally, we aggregated results across topics to examine overall topic-level wins. In the observed data, Democrats were the majority winner in 64%

of topics based on argument convincingness (see subsection 5.4). Randomization testing showed that, despite near equivalence in global pairwise wins, Democrats won significantly more topics than expected by chance. This indicates a topic-level advantage for Democrats that was not evident from global pairwise comparisons alone.

6 Conclusion

We presented a computational analysis of campaign speeches, focusing on emotion's role in persuasive discourse. To achieve this, we trained models for argument detection and argument quality assessment. Our argument detection model achieved **0.87** precision, accurately identifying argumentative statements in political speech. The argument quality model achieved **0.78** accuracy, offering reliable persuasiveness evaluation. Using these models, we found that both parties rely heavily on argumentation in their U.S. political speeches. To assess argument effectiveness, we applied a rhetorical lens and found that high-quality arguments boost persuasion, with Democrats winning 65% of the time. Our analysis also shows that politicians strategically use emotional cues to enhance message persuasiveness. While our study is purely computational, we view human supervision as an important direction for future work to validate and complement these findings. Future work will analyze arguments logically to identify fallacies in political speech. Combining logical, rhetorical, and emotional analyses could reveal whether fallacious arguments are especially persuasive and which emotions they evoke.

Limitations

There are few limitations to our study that should be noted. First, while training the argument detection model, we did not split the data based on topics. Although we subsequently tested the model using topic-based splits and observed minimal differences in performance, this approach may still overlook subtle topic-specific patterns in argumentation. Second, our emotion detection model assigns a single dominant emotion to each argument. In reality, multiple emotions with varying intensities can be associated with a single utterance. While we partially addressed this by aggregating emotion distributions at the speaker and topic levels, a more fine-grained analysis would require a system capable of capturing the full emotion distribution for each individual argument. Third, to identify argumentative content in political speeches, we segment the text into non-overlapping chunks of up to 512 tokens. While this chunking strategy allows us to capture arguments that span multiple sentences, it may also dilute argumentative signals, potentially causing some argumentative content to be misclassified as non-argumentative. Future work could explore sequence-level or hierarchical modeling approaches to more precisely capture argumentative structure within longer text spans. Finally, we acknowledge that pretrained language models may encode political or ideological biases, which could influence convincingness predictions. Although our randomization-based analyses mitigate some aggregate effects, we cannot fully rule out such biases, and we view this as an important limitation of the current study.

References

Segun T Aroyehun, Almog Simchon, Fabio Carrella, Jana Lasser, Stephan Lewandowsky, and David Garcia. 2025. Computational analysis of us congressional speeches reveals a shift from evidence to intuition. *Nature Human Behaviour*, pages 1–12.

Mohamed S Benlamine, Serena Villata, Ramla Ghali, Claude Frasson, Fabien Gandon, and Elena Cabrio. 2017. Persuasive argumentation and emotions: An empirical evaluation with users. In *International Conference on Human-Computer Interaction*, pages 659–671. Springer.

Sahbi Benlamine, Maher Chaouachi, Serena Villata, Elena Cabrio, Claude Frasson, and Fabien Gandon. 2015. Emotions in argumentation: an empirical evaluation. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015, Buenos Aires, Argentina, July 25-31, 2015*, pages 156–163. AAAI Press.

Luca Beurer-Kellner, Marc Fischer, and Martin Vechev. 2023. Prompting is programming: A query language for large language models. *Proceedings of the ACM on Programming Languages*, 7(PLDI):1946–1969.

Amparo Elizabeth Cano-Basave and Yulan He. 2016. A study of the impact of persuasive argumentation in political debates. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1405–1413, San Diego, California. Association for Computational Linguistics.

Ioannis Chalkiadakis, Louise Anglès d'Auriac, Gareth W Peters, and Divina Frau-Meigs. 2025. A text dataset of campaign speeches of the main tickets in the 2020 us presidential election. *Scientific Data*, 12(1):662.

Dennis Chong and James N Druckman. 2007. A theory of framing and opinion formation in competitive elite environments. *Journal of communication*, 57(1):99–118.

Alan Darmasaputra Chowanda, Albert Richard Sanyoto, Derwin Suhartono, and Criscentia Jessica Setiadi. 2017. Automatic debate text summarization in online debate forum. *Procedia computer science*, 116:11–19.

James N Druckman. 2022. A framework for the study of persuasion. *Annual Review of Political Science*, 25:65–88.

Barbara Early. 2015. The righteous mind: Why good people are divided by politics and religion, by jonathan haidt: (2012). new york, ny: Pantheon books, illustrated, 419 pp., $28.95 (hard cover).

Liat Ein-Dor, Eyal Shnarch, Lena Dankin, Alon Halfon, Benjamin Sznajder, Ariel Gera, Carlos Alzate, Martin Gleize, Leshem Choshen, Yufang Hou, Yonatan Bilu, Ranit Aharonov, and Noam Slonim. 2020. Corpus wide argument mining - A working solution. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 7683–7691. AAAI Press.

Kevin Finity, Ramit Garg, and Max McGaw. 2021. A text analysis of the 2020 us presidential election campaign speeches. In *2021 Systems and Information Engineering Design Symposium (SIEDS)*, pages 1–6. IEEE.

Gloria Gennaro and Elliott Ash. 2022. Emotion and reason in political language. *The Economic Journal*, 132(643):1037–1059.

Martin Gleize, Eyal Shnarch, Leshem Choshen, Lena Dankin, Guy Moshkowich, Ranit Aharonov, and Noam Slonim. 2019. Are you convinced? choosing the more convincing evidence with a Siamese network. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 967–976, Florence, Italy. Association for Computational Linguistics.

Lynn Greschner and Roman Klinger. 2024. Fearful falcons and angry llamas: Emotion category annotations of arguments by humans and llms. *ArXiv preprint*, abs/2412.15993.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *ArXiv preprint*, abs/2203.05794.

David J Grüning and Thomas W Schubert. 2022. Emotional campaigning in politics: being moved and anger in political ads motivate to support candidate and party. *Frontiers in Psychology*, 12:781851.

Ivan Habernal and Iryna Gurevych. 2016. What makes a convincing argument? empirical analysis and detecting attributes of convincingness in web argumentation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1214–1223, Austin, Texas. Association for Computational Linguistics.

Shohreh Haddadan, Elena Cabrio, and Serena Villata. 2019. Yes, we can! mining arguments in 50 years of US presidential campaign debates. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4684–4690, Florence, Italy. Association for Computational Linguistics.

Hans JG Hassell, Christopher D Johnston, Jessica Khan, and Edgar Cook. 2022. The nature and impact of emotional content in congressional candidate emails to supporters. *Electoral Studies*, 79:102501.

Rositsa Ivanova, Thomas Huber, and Christina Niklaus. 2024. Let's discuss! quality dimensions and annotated datasets for computational argument quality assessment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20749–20779.

John Lawrence and Chris Reed. 2019. Argument mining: A survey. *Computational Linguistics*, 45(4):765–818.

Jennifer S Lerner, Ye Li, Piercarlo Valdesolo, and Karim S Kassam. 2015. Emotion and decision making. *Annual review of psychology*, 66(1):799–823.

Ran Levy, Ben Bogin, Shai Gretz, Ranit Aharonov, and Noam Slonim. 2018. Towards an argumentative content search engine using weak supervision. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2066–2081, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Marco Lippi and Paolo Torroni. 2016. Argument mining from speech: Detecting claims in political debates. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 2979–2985. AAAI Press.

Mounika Marreddy, Subba Reddy Oota, Venkata Charan Chinni, Manish Gupta, and Lucie Flek. 2025. Usdc: A dataset of user stance and dogmatism in long conversations. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23715–23759.

Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. 2018. Never retreat, never retract: Argumentation analysis for political speeches. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 4889–4896. AAAI Press.

Sovesh Mohapatra and Somesh Mohapatra. 2022. Sentiment is all you need to win US presidential elections. In *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, pages 15–20, Taipei, Taiwan. Association for Computational Linguistics.

Richard E Petty and John T Cacioppo. 2012. *Communication and persuasion: Central and peripheral routes to attitude change*. Springer Science & Business Media.

Trong Huy Phan and Kazuma Yamamoto. 2020. Resolving class imbalance in object detection with weighted cross entropy losses. *ArXiv preprint*, abs/2006.01413.

Minghui Qiu, Yanchuan Sim, Noah A. Smith, and Jing Jiang. 2015. Modeling user arguments, interactions, and attributes for stance prediction in online debate forums. In *Proceedings of the 2015 SIAM International Conference on Data Mining, Vancouver, BC, Canada, April 30 - May 2, 2015*, pages 855–863. SIAM.

Carlotta Quensel, Neele Falk, and Gabriella Lapesa. 2025. Investigating subjective factors of argument strength: Storytelling, emotions, and hedging. In *Proceedings of the 12th Argument mining Workshop*, pages 126–139.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2024. Qwen2.5 technical report.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Valerie Rodriguez-Hernandez, Vanesa Hidalgo, and Alicia Salvador. 2024. Emotional and cognitive processes underlying persuasion, moderating factors, and physiological reactions: A systematic review. *Psychological Reports*, page 00332941241291497.

Sandy Ross. 2006. Don't think of an elephant: Know your values and frame the debate. *Melbourne Journal of Politics*, 31:145–149.

Tamanna M Shah. 2022. Emotions in politics: A review of contemporary perspectives and trends. *International Political Science Abstracts*, 74(1):1–14.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th International Conference on World Wide Web, WWW 2016, Montreal, Canada, April 11 - 15, 2016*, pages 613–624. ACM.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, and 1 others. 2025. Gemma 3 technical report. *ArXiv preprint*, abs/2503.19786.

Henning Wachsmuth, Khalid Al-Khatib, and Benno Stein. 2016. Using argument mining to assess the argumentation quality of essays. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1680–1691, Osaka, Japan. The COLING 2016 Organizing Committee.

Overview of Appendix Sections

- Section A: Dataset Statistics
- Section B: LLM Prompts
- Section C: CSTD Topics

A Dataset Statistics

Statistic	Train	Validation
Total instances	25,542	6,386
Unique topics	240	236
Avg. sentences per topic	106.4	27.1
Avg. sentence length (tokens)	31.0	31.0

Table 2: Dataset statistics for the ARGDETECT corpus split into train and validation.

Statistic	Train	Validation
Total instances	4,319	1,378
Unique topics	48	21
Avg. instances per topic	90	65
Avg. sentence length (tokens)	28	28

Table 3: Dataset statistics for the EVICONV corpus across train and validation splits.

In this section, we present the full statistics of the dataset used for argument detection (Table 2) and argument quality assessment (Table 3) training.

B LLM Prompts

B.1 Argument Detection

Figure 5 shows the prompt that we used as annotation guidelines to obtain argument detection annotations from LLMs.

B.2 Argument Quality Prompt

Figure 6 shows the prompt that we used as annotation guidelines to obtain argument quality assessment annotations from LLMs.

C Topics in CSTD

88 topics are shown in Table 4 extracted using BERTopic and labelled using LLMs like Gemma3, Qwen2.5 and Llama3.2.

C.1 Top 10 topics

Table 5 shows top 10 topics for Democratic party where they exhibit higher convincingness. Similarly, Table 6 shows the same for the Republican party.

Argument Detection

You are an expert in argument mining and discourse analysis. Your task is to determine if a given Sentence/Statement functions as an argument (a claim, reason, or piece of evidence) for or against the provided Topic.

The output should be a single integer:

- Label: 1 if the Sentence/Statement acts as an argument (pro or con) for the Topic.
- Label: 0 if the Sentence/Statement is purely descriptive, definitional, or factually related to the Topic but does not attempt to persuade, support, or challenge a viewpoint about it.

In other words, check if the Sentence/Statement offers a reason why someone should think a certain way about the Topic.

Example 1:
Topic: Amnesty International
Sentence: In a report of November 1988, Amnesty International said that the number of prisons had increased to 644 and their capacity had been raised from 55,000 to more than 80,000 [REF].
Label: 0

Example 2:
Topic: Environmental technology
Sentence: Designed to transform the site into a 21st-century global laboratory which brings together business, education and green technologies, it aims to create over 1,000 jobs for the local area and generate £29m private sector investment.
Label: 1

Now answer the following:

```
Topic: {topic}
Sentence: {sentence}
Label:
```

Figure 5: Prompt for Argument Detection

Argument Quality Assessment

You are an expert at evaluating the quality of arguments based on evidence provided. Given a claim and two pieces of evidence, your task is to determine which piece of evidence better supports the claim.

Even if both pieces of evidence are weak or imperfect, you must choose the one that is relatively better at supporting the claim.

Example 1:
Claim: We should legalize prostitution
Evidence 1: The appellants' argument then, more precisely stated, is that in criminalizing so many activities surrounding the act itself, Parliament has made prostitution de facto illegal if not de jure illegal.
Evidence 2: Feminists who hold such views on prostitution include Kathleen Barry, Melissa Farley, Julie Bindel, Sheila Jeffreys, Catharine MacKinnon and Laura Lederer; the European Women's Lobby has also condemned prostitution as "an intolerable form of male violence".
Answer: 1

Example 2:
Claim: We should subsidize condoms
Evidence 1: In 2009, Lewis strongly criticized Pope Benedict XVI's assertion that condom use only makes the AIDS crisis worse.
Evidence 2: Green said that according to the "best studies," condoms makes people take wilder sexual risks, thus worsening the spread of the disease.
Answer: 2

Now answer the following:

```
Claim: {claim}
Evidence 1: {evidence_1}
Evidence 2: {evidence_2}
Answer:
```

Figure 6: Prompt for Argument Quality Assessment

ID	Topic	ID	Topic
1	Veterans Healthcare Reform & Recognition	45	Venezuelan Democratic Crisis (Regime Transition)
2	U.S. Military Modernization Budget	46	Pennsylvania Fracking Regulations
3	Border Control Policies	47	Job Dignity & Workplace Respect
4	Georgia Senate Runoff Elections (Perdue–Warnock, Loeffler–Ossoff)	48	U.S. New England Primaries
5	Jerusalem & Golan Heights Sovereignty Recognition	49	US Election Racial Tensions: Kamala Harris's Controversies
6	ISIS Caliphate Collapse	50	Voting Participation & Early Voting
7	Iran Nuclear Sanctions	51	Voter Integrity & 2020 US Election Contests
8	Mass Shootings at Houses of Worship (Anti-Semitism)	52	NATO Alliance Funding Obligations
9	U.S. Space Governance and NASA Growth	53	Middle Class Empowerment & Labor Strength
10	Lunar and Martian Space Exploration	54	Progressive Tax Reforms
11	Small Business Policy and Political Alliances	55	Federal Agency Position Classification & Executive Accountability
12	Tax Cuts & Deregulation	56	Universal Access to Comprehensive Healthcare (Including Pre-existing Conditions)
13	Minority Unemployment and Wage Growth	57	White Supremacist Violence (Charlottesville Incident)
14	Vocational Training & Apprenticeship Initiative (for U.S. Workers)	58	Climate Resilient Clean Energy Infrastructure
15	Criminal Justice Reform & Sentencing Equity	59	Racial Equity and Community Policy
16	Gun Control Regulation	60	Foreign Election Interference Risk to U.S. Democracy
17	National Prayer and Faith Holiday	61	Refugee Policy & Admissions (Resettlement Increase)
18	American Values Advocacy: Constitution, Law Enforcement, Faith, Family, History	62	Empowering Women Against Gender-Based Violence
19	U.S. Agricultural Policy & Farmer Support	63	Policy on National Security, Gun Control, & Immigration Economics
20	USMCA's Effect on American Jobs and Agriculture	64	North Korea's Nuclear Diplomacy (U.S., Russia focus)
21	Energy Independence & Resource Dominance	65	U.S. Intervention in Syria–Turkey Kurdish Conflict
22	U.S.-Iowa Economic Impact of USMCA	66	Suburban Women's & Housing Safety Regulations
23	National Defense Expansion	67	Historic, Diverse, Inclusive Cabinet – Environmental Priorities
24	Universal Education Equity	68	US COVID-19 Health Policy & Treatments
25	Supreme Court Appointments (Pro-Life Focus)	69	Affordable Prescription Drugs & Transparent Healthcare
26	Voting Rights and Civil Rights Advocacy	70	City Crime and Civil Unrest
27	Manufacturing Economic Renaissance	71	Dreamer Citizenship Policy
28	Impact of Soleimani's Death on Terrorism in Iraq	72	Hispanic Voting Trends in Florida
29	Ukraine Impeachment Inquiry (Russia Ties)	73	Right-to-Try Drug Access Laws
30	Michigan Automotive Resurgence (Political Implications)	74	COVID-19 Mask Policies & Public Health
31	Late-Term Abortion and Fetal Rights	75	Mail-In Ballot Integrity and Fraud Issues
32	Economic Freedom and Anti-Socialism in America	76	America Reclamation Movement
33	Police Reform & Racial Equity	77	Public Gathering Management Policies & Controversies
34	Law Enforcement Reform & Support	78	Stock Market Performance & Investing
35	COVID-19 Manufacturing Response (Ventilators)	79	Tax Relief for Family Farms and Small Businesses
36	COVID-19 Testing Capacity Growth (US, Feb '20)	80	Opioid Crisis Treatment Across Regions (Montana & Kentucky)
37	Pandemic Economy Aid for Small Businesses & Workers	81	Politics of Anti-Immigration in Minnesota (re: Omar)
38	COVID-19 Mishandling (Early 2020)	82	Household Appliance Water Efficiency Regulations
39	Veterans Healthcare Accountability Act	83	Media Ownership Impact on News Reporting & Bias
40	Police Funding Reform & Systemic Racism	84	Boeing Contract Cancellation Fees
41	Global COVID-19 Vaccination Efforts	85	Monument Preservation vs Removal Conflict
42	U.S. Election: Allegations of Government Corruption and Family Scandals	86	Politics and Law Enforcement
43	Firearms Regulation and Second Amendment	87	Rape Kit Backlog & Victim Support Services
44	Border Security for Refugees from Conflict Areas	88	September 11 Heroes & Flight 93 Tribute

Table 4: Topics extracted using BERTopic and labelled using LLMs like Gemma3, Qwen2.5 and Llama3.2

Topic	Dem Wins	Rep Wins	Dem Win Rate
COVID-19 Mask Policies & Public Health	54	0	1.000
Border Security for Refugees from Conflict Areas	30	0	1.000
City Crime and Civil Unrest	34	0	1.000
Job Dignity & Workplace Respect	74	0	1.000
Michigan Automotive Resurgence	27	0	1.000
Pennsylvania Fracking Regulations	72	6	0.923
Hispanic Voting Trends in Florida	11	1	0.917
USMCA (Iowa Economic Impact)	56	7	0.889
September 11 Heroes & Flight 93 Tribute	14	2	0.875
Voting Participation & Early Voting	47	7	0.870

Table 5: Top 10 topics where Democratic speakers were more convincing.

Topic	Dem Wins	Rep Wins	Dem Win Rate
U.S. New England Primaries	0	4	0.000
America Reclamation Movement	0	7	0.000
White Supremacist Violence (Charlottesville)	1	42	0.023
Minority Unemployment & Wage Growth	16	124	0.114
Lunar and Martian Space Exploration	11	65	0.145
Iran Nuclear Sanctions	15	54	0.217
Empowering Women Against Gender-Based Violence	18	48	0.273
Veterans Healthcare Reform & Recognition	55	141	0.281
Mail-In Ballot Integrity & Fraud Issues	309	559	0.356
NATO Funding Obligations	36	58	0.383

Table 6: Top 10 topics where Republican speakers were more convincing.