From Debates to Diplomacy: Argument Mining Across Political Registers

Maria Poiaganova
Marketing for Social Impact
University of Zürich
Zürich, Switzerland
maria.poiaganova@business.uzh.ch

Manfred Stede
Applied Computational Linguistics
University of Potsdam
Potsdam, Germany
stede@uni-potsdam.de

Abstract

This paper addresses the problem of cross-register generalization in argument mining within political discourse. We examine whether models trained on adversarial, spontaneous U.S. presidential debates can generalize to the more diplomatic and prepared register of UN Security Council (UNSC) speeches. To this end, we conduct a comprehensive evaluation across four core argument mining tasks. Our experiments show that the tasks of detecting and classifying argumentative units transfer well across registers, while identifying and labeling argumentative relations remains notably challenging, likely due to register-specific differences in how argumentative relations are structured and expressed. As part of this work, we introduce *ArgUNSC*, a new corpus of 144 UNSC speeches manually annotated with claims, premises, and their argumentative links. It provides a resource for future in- and cross-domain studies and novel research directions at the intersection of argument mining and political science.

1 Introduction

Argumentation is integral to human communication, enabling individuals to express opinions, persuade others, and collaboratively reason about the world. As artificial intelligence systems increasingly assist humans, both in everyday interactions and in high-stakes decision-making scenarios, their ability to detect and interpret arguments is more critical than ever. Therefore, Argumentation Mining (AM) plays a central role in such systems, enabling them to identify and structure argumentative content across a variety of domains, spanning legal decision support (Habernal et al., 2024), educational tools for developing students' reasoning skills (Wambsganss et al., 2021), social media analysis (Feger and Dietze, 2024; Chakrabarty et al., 2019), and even autonomous debating technologies (Slonim et al., 2021).

Building robust AM systems is tightly connected to high-quality annotated data. However, creating such datasets across all potential domains and contexts is time-consuming, costly, and intellectually demanding. To address this challenge, *cross-domain* generalization—a strategy where models trained in one domain (e.g., legal) are evaluated in another (e.g., medical)—has emerged (Daxenberger et al., 2017; Schaefer et al., 2022; Gemechu et al., 2024). At the same time, relatively little attention has been paid to *cross-register* generalization—a special case of domain transfer where the broader discourse remains consistent but the rhetorical style, structure, or communicative setting varies. This scenario appears promising and challenging at the same time, as, on the one hand, registers within the same domain often share core argumentative structures, allowing for potential knowledge transfer, particularly when communicative goals such as persuasion or justification are preserved. On the other hand, even subtle differences in style, lexical choices, or discourse organization can hinder generalization.

To address this open question, our paper focuses on the challenge of cross-register generalization in AM within the domain of political discourse. We contrast U.S. presidential debates and United Nations Security Council (UNSC) speeches, which represent markedly different registers. Presidential debate discourse is often spontaneous, and aimed at persuading a public audience. In contrast, UNSC speeches are mostly prepared, and delivered in formal institutional settings to articulate national positions.

Our goal is to investigate whether argumentation models trained on political speech of one register can generalize to a speech with a different register. We evaluate this across four core AM tasks: (1) Argumentative Component Segmentation (ACS) – detecting argumentative components (*claims* and *premises*); (2) Argumentative Component Classifi-

Proceedings of the 12th Argument Mining Workshop, pages 205–216
July 31, 2025 ©2025 Association for Computational Linguistics

cation (ACC) – distinguishing between claims and premises; (3) Argumentative Relation Identification (ARI) – determining whether a claim and a premise are argumentatively related; and (4) Argumentative Relation Classification (ARC) – identifying whether the relation is *support* or *attack*. For language modeling, we use encoder-based architectures (BERT and RoBERTa) and evaluate performance in both in-register and cross-register settings. Additionally, we prompt GPT-4 in zero- and few-shot setups and compare its performance to that of fine-tuned models.

Beyond this systematic cross-task and cross-model evaluation, a major part of our contribution lies in releasing a novel corpus of 144 UNSC speeches, annotated with claims, premises, and the relations between them.

Our results reveal that ACS and ACC tasks generalize well across registers, whereas ARI and ARC do not, highlighting the greater complexity of relation-level tasks and their sensitivity to register variation. Additionally, LLMs consistently underperform compared to encoder models fine-tuned both in- and cross-register scenarios, with particularly large performance gaps on ACS and ACC tasks.

2 Related Work

2.1 Political Argument Mining

Our work contributes to the literature on political argument mining, motivating a review of existing political corpora and the specific AM tasks they support. For example, Menini et al. (2018) introduce a corpus of 1,462 manually annotated argument pairs drawn from Nixon and Kennedy's 1960 presidential campaign speeches. The pairs are labeled with support and attack relations across five major political topics.

Similarly, Visser et al. (2020) present the $US2016$ corpus, which includes transcriptions of televised debates leading up to the 2016 US presidential election, as well as audience reactions collected from Reddit.[1]

Lippi and Torroni (2016a) compile an original dataset based on the 2015 UK political election debates, combining textual and audio features and test whether spoken language cues improve claim detection.

Another multimodal corpus is presented by Mestre et al. (2021). Their *M-Arg* dataset is based on the US 2020 presidential debates and includes both audio and transcripts, annotated with claims and premises and the argumentative relationship between them across 4,104 sentence pairs.

Haddadan et al. (2019) present a large-scale corpus of 39 U.S. presidential debates spanning from 1960 to 2016, annotated with claims and premises. They further explore argument filtering and argument component classification. A recent extension enriches the corpus with relation annotations and labels for argumentative fallacy types (Goffredo et al., 2022). We select this corpus to represent the presidential debates register in our study.

2.2 Argument Mining under Low-Resource Conditions

2.2.1 Cross-Domain Generalization

The challenge of transferring models across domains or text genres has been widely studied in NLP more broadly (Hupkes et al., 2023), and remains particularly difficult in the context of AM. In an early study, Ajjour et al. (2017) demonstrate significant generalizability issues for the argument unit segmentation task across three datasets. Daxenberger et al. (2017) undertake systematic experiments in cross-domain claim classification in six different datasets, and find generally high degradation compared to in-domain performance. Using qualitative analysis, they show that the underlying notions of claim in the datasets vary significantly. Similarly, Schaefer et al. (2022) use four corpora of varying genres and sizes and conclude that large training sets, homogeneous claim ratios, and less formal language tend to improve generalization. In a series of relation identification tasks, Gemechu et al. (2024) propose a benchmark architecture encompassing three approaches and conduct experiments on—and across—eight datasets. In line with previous cross-domain studies, they observe consistently poor performance when detecting support and attack relations in corpora unseen during training.

Turning from cross-domain to cross-register setups closer to ours, Blokker et al. (2020) examine the generalizability of claim detection models by training on newspaper data and testing on political party manifestos. Despite linguistic and conceptual differences between formats, their BERT-based model shows strong cross-text performance and strong overlap in party positions across registers.

[1] https://www.reddit.com/

2.2.2 Large Language Models in AM

A rapidly growing body of literature highlights the remarkable capabilities of Large Language Models (LLMs) in argument mining (Chen et al., 2024; Favero et al., 2025; Cabessa et al., 2025; Sviridova et al., 2024). LLMs are particularly well-suited for low-resource settings, showing strong performance even with simple instruction prompts. However, given their recent emergence, research remains limited and evidence mixed regarding their performance on AM tasks compared to other state-of-the-art models. For instance, Gorur et al. (2025) examine argument relation identification and find that prompted LLMs significantly outperform RoBERTa across 11 datasets. On the other hand, Ruiz-Dolz and Lawrence (2023) find that fine-tuned RoBERTa outperforms GPT-4 in most cases in the context of argumentative fallacy detection.

3 Data

3.1 US Presidential Debates

As a starting point for our register transfer experiments, we focus on presidential debates discourse. We adopt the large-scale US-ElecDeb16To60 v.01 corpus (hereafter, USElecDeb), introduced by Haddadan et al. (2019). The corpus comprises transcripts from 39 U.S. presidential and vice-presidential debates spanning from 1960 to 2016. These transcripts were originally obtained from the Commission on Presidential Debates[2].

The USElecDeb corpus contains annotations of argumentative components, namely claims and premises. According to Haddadan et al. (2019), in political debate discourse, a claim may take the form of an advocated policy, a candidate's stance on a policy, an opinion on a particular issue, or their personal judgment. To justify their claims, politicians provide premises (sometimes referred to as evidence in the AM literature (e.g., Cheng et al. (2022); Lippi and Torroni (2016b)), which may include references to specific events, data, outcomes of past policies, etc.

Importantly, annotations in the original corpus are made on the component level, with components defined as the minimal discourse units that independently convey argumentative meaning. Such units can span the entire sentence or be more granular, e.g., take the form of a clause. For modeling purposes, the authors map the component-level annotations to the sentence level, a setup we adopt in our experiments as well. Table 1 provides the sentence-level distribution of claims and premises in the USElecDeb corpus.

Level	Total	Arg	Non-Arg	Claim	Premise
Sent.	29.621	22.280	7.252	11.964	10.316

Table 1: Distribution of argumentative sentence types in the USElecDeb.

While the original dataset did not include relation annotations, these were later introduced by Goffredo et al. (2022) as part of a study on fallacy detection. In addition to augmenting the corpus with relational links (support or attack) between components, this extended version also includes transcripts from Biden-Trump debates held in 2020.[3] We use this enhanced version for our experiments on Argumentative Relation Identification (ARI) and Argumentative Relation Classification (ARC). The summary statistics on support vs. attack sentence pairs is presented in Table 2.

Level	Total	Support	Attack
Sent.	25.524	21.689	3.835

Table 2: Distribution of support/attack sentence-pairs in the USElecDeb.

Example (1) represents an argumentative structure in USElecDeb. Claims are marked in **bold**, premises in *Italics*, and the component boundaries are additionally indicated by [square brackets]. In this example, both premises support the claim.

(1) Nixon-Kennedy, September 26, 1960:

NIXON: We often hear gross national product discussed, and in that respect may I say that [*when we compare the growth in this Administration with that of the previous Administration that then there was a total growth of eleven percent over seven years*]$_{\text{Premise}_1}$; [*in this Administration there has been a total growth of nineteen percent over seven years*]$_{\text{Premise}_2}$. **[That shows that there's been more growth in this Administration than in its predecessor]**$_{\text{Claim}}$.

[2] https://www.debates.org/

[3] https://github.com/pierpaologoffredo/ElecDeb60to20/tree/main

3.2 UN Security Council Speeches

The United Nations Security Council (UNSC) is a principal body responsible for maintaining international peace and security; it convenes when global conflicts, crises, or threats to peace require collective diplomatic response. The UNSC discourse was selected as a contrasting register to presidential debates in our cross-register experiments. Its largely formal and pre-written language differs markedly from the spontaneous and often emotionally charged language of debates. Beyond this stylistic divergence, it also holds intrinsic value for argument mining due to its high-stakes discourse in which nations articulate their positions through structured and strategic reasoning.

To collect the data, we use the raw corpus of UNSC speeches published by Schönfeld et al. (2019). We select speeches from the years 2014 to 2018, a period marked by the onset of the Russia-Ukraine conflict—a topic that prompted diverse and rich argumentative positions from various countries. In addition to discussions of this conflict, a few speeches address issues related to the UNSC's Women, Peace, and Security (WPS) agenda. The final dataset includes 144 speeches delivered by representatives from 24 different nations. Appendix A details the distribution of speeches by country and year. Notably, our corpus was developed in parallel with UNSCon (Zaczynska et al., 2024) and contains 44 overlapping speeches, enabling joint analyses of argumentation structures and conflict discourse in diplomatic setting in future work.

Level	Total	Arg	Non-Arg	Claim	Premise
Sent.	4.765	4.105	660	2.081	2.024

Table 3: Distribution of argumentative types across sentences in the ArgUNSC.

Level	Total	Claim	Premise
Component	4.584	2.328	2.256

Table 4: Distribution of claim and premise components in the ArgUNSC.

During annotation, claims and premises were marked on a component level, following Haddadan et al. (2019). To identify and distinguish argument components, we initially relied on the guidelines provided by the authors of USElecDeb. We note that, as the genres are slightly different, we met several types of arguments that are specific to our data, which resulted in some annotation guideline extensions. In particular, diplomatic speeches focusing on military conflict often include *claims* that express the speaker nation's interpretation or evaluation of the current situation, their position on the actions of other parties, or proposals for conflict mitigation. Typical *premises* in this context involve references to concrete events or official documents. These statements frequently include details such as dates, actors involved, actions taken, and consequences observed, as illustrated in Example (2).

Level	Total	Support	Attack
Component	2.973	2.623	350

Table 5: Distribution of support/attack argumentative-component pairs in the ArgUNSC.

N premises	N components
no premise	640
one premise	1007
two premises	381
three premises	156
> three premises	144

Table 6: Distribution of the number of premises per one claim in the ArgUNSC corpus.

A *support* relation indicates that the premise provides a reason to believe the claim, as in Example (2), while an *attack* relation represents an opposing position—typically anticipating or addressing potential objections a hearer might raise, as shown in Example (3).

(2) United Kingdom, 2014:

[The situation in eastern Ukraine has continued to deteriorate]$_{\text{Claim}}$. [*Armed groups stormed the Prosecutor's office in Donetsk yesterday, further increasing the number of Government buildings occupied since the 17 April Geneva agreement*]$_{\text{Premise}}$

(3) China, 2014:

China notes that, [*since the signing of the Minsk agreements between the Ukrainian Government and eastern militias at the beginning of September, there have been no large-scale armed clashes in eastern Ukraine*]$_{\text{Premise}}$. However, **[the security situation on the ground still remains fragile with sporadic violent attacks in violation of the ceasefire agreement, causing casualties and damage**

to infrastructure]$_{\text{Claim}}$

We report statistics of the dataset for both sentence- and component levels. (Tables 3 and 4). Like in the USElecDeb corpus, we observe that claims slightly outnumber premises, which is not rare in political discourse, where speakers do not always provide premises to justify their claims.

We also note that, according to our guidelines, claims can relate to more than one premise at a time. Similarly, one premise may relate to one or more claims. In our corpus, we observe considerable variation in the number of premises per claim, ranging from none to more than three, as shown in Table 6. Regarding relations, as seen from the tables 2 and 5, in both UNSC speeches and presidential debates, premises predominantly support rather than attack claims, reflecting speakers' tendency to reinforce their position, no matter if one is speaking on behalf of a country or campaigning for the presidency.

Three annotators with backgrounds in computational linguistics participated in the annotation process. First, A1 and A2 collaboratively developed the annotation guidelines, using several test speeches to explore the intricacies of the corpus and iteratively refine the guidelines. After this pilot phase, A1 completed the full annotation of the corpus. Subsequently, A2 independently annotated 29 documents (excluding the test set), representing 20% of the corpus), labeling argumentative components as claims or premises. The annotation of argumentative relations (support/attack) was then performed by a third annotator (A3), who had access to the existing fixed component boundaries established by A2. The annotation process was carried out using the INCEpTION software (Klie et al., 2018).

Inter-annotator agreement (IAA) was measured using Cohen's κ statistic, calculated at the sentence level. First, we assessed whether annotators agreed on the sentence's argumentative status ($\kappa = 0.69$). Next, considering only sentences both annotators identified as argumentative, we measured agreement on whether the sentence contained a claim or a premise ($\kappa = 0.77$). To compute IAA for relations, we considered the sentence-pair level. Within each speech, we generated the set of all possible claim-premise pairs of sentences and calculate agreement on whether each pair is labeled as support, attack, or no relation ($\kappa = 0.68$). Thus, we report overall *substantial* agreement on argument component and relation annotation tasks (Artstein and Poesio, 2008). A1's labels serve as gold standard.

4 Methodology

4.1 Argument Mining Pipeline

Following Liu et al. (2023), we divide argumentation mining into the following four steps. We approach each step as a binary classification task, precisely formulated as follows.

Argumentative Component Segmentation (ACS). Given a sentence X, predict whether it *contains* an argumentative component (can be either a claim or a premise) or not.

Argumentative Component Classification (ACC). Given an argumentative sentence X, predict whether it *contains* a claim or a premise.

Argumentative Relation Identification (ARI). Given a pair of sentences (or components) (X, Y), the task is to predict whether X is argumentatively related to Y (as either support or attack), or not. For training, we randomly generate an equal number of unrelated pairs by sampling sentences (or components) from speeches in close temporal proximity—specifically, within eight speeches before or after the given speech—thereby ensuring comparable contextual conditions.

Argumentative Relation Classification (ARC). Given a pair of argumentatively related sentences *(or components)* (X, Y), predict whether X and Y are in a support or attack relationship.

All tasks are first performed at the sentence level. For ACS, a sentence is labeled argumentative if it contains at least one argumentative component—claim or premise. For ACC, since a sentence may contain both a claim and a premise, we follow Haddadan et al. (2019) and assign the label based on the longer component. For ARI and ARC, we consider a sentence pair (X, Y) as related if a component in X is linked to a component in Y. While the sentence-level setup is straightforward, we acknowledge that it may obscure information when multiple components appear in the same sentence. In our corpus, this happens in about 7% of cases, which poses particular challenges for relation-based tasks. Therefore, we also report component-level results for ARI and ARC.

4.2 Experimental Setup

We treat all tasks as sequence classification and fine-tune transformer-based encoders using the *bert-for-*

Evaluation	Label	Majority Vote	BERT			RoBERTa		
		F1	P	R	F1	P	R	F1
IR–USElecDeb$_{\text{sentence}}$	Argument	0.810	0.877	0.939	0.907	0.883	0.945	0.913
	Not Argument	0.000	0.717	0.541	0.617	0.745	0.563	0.641
	Avg Macro	0.551	0.797	0.740	0.762	0.814	0.754	0.777
IR–ArgUNSC$_{\text{sentence}}$	Argument	0.926	0.936 ± 0.008	0.973 ± 0.008	0.954 ± 0.004	0.937 ± 0.008	0.978 ± 0.005	0.957 ± 0.004
	Not Argument	0.000	0.778 ± 0.043	0.586 ± 0.055	0.667 ± 0.038	0.810 ± 0.029	0.591 ± 0.056	0.682 ± 0.039
	Avg Macro	0.463	0.857 ± 0.022	0.780 ± 0.026	0.810 ± 0.021	0.874 ± 0.015	0.784 ± 0.027	0.819 ± 0.021
CR$_{\text{sentence}}$	Argument	0.926	0.916	0.980	0.947	0.930	0.965	0.947
	Not Argument	0.000	0.776	0.441	0.562	0.713	0.547	0.619
	Avg Macro	0.463	0.846	0.710	0.754	0.822	0.756	0.783

Table 7: F1-score for the majority vote baseline and Precision (P), Recall (R), and F1-scores for BERT and RoBERTa. Task: **Argumentative Component Segmentation (ACS)** in in-register and cross-register settings.

Evaluation	Label	Majority Vote	BERT			RoBERTa		
		F1	P	R	F1	P	R	F1
IR–USElecDeb$_{\text{sentence}}$	Claim	0.677	0.671	0.754	0.710	0.660	0.806	0.726
	Premise	0.000	0.705	0.614	0.656	0.736	0.567	0.640
	Avg Weighted	0.346	0.688	0.685	0.684	0.698	0.689	0.684
IR–ArgUNSC$_{\text{sentence}}$	Claim	0.673	0.716 ± 0.006	0.850 ± 0.019	0.777 ± 0.006	0.757 ± 0.029	0.766 ± 0.027	0.761 ± 0.009
	Premise	0.000	0.809 ± 0.014	0.653 ± 0.028	0.722 ± 0.014	0.756 ± 0.014	0.745 ± 0.045	0.750 ± 0.020
	Avg Weighted	0.341	0.762 ± 0.006	0.753 ± 0.007	0.750 ± 0.008	0.757 ± 0.014	0.755 ± 0.014	0.755 ± 0.013
CR$_{\text{sentence}}$	Claim	0.673	0.698	0.801	0.746	0.719	0.772	0.745
	Premise	0.000	0.759	0.643	0.696	0.747	0.690	0.717
	Avg Weighted	0.341	0.728	0.723	0.721	0.733	0.732	0.731

Table 8: F1-score for the majority vote baseline and Precision (P), Recall (R), and F1-scores for BERT and RoBERTa. Task: **Argumentative Component Classification (ACC)** in in-register and cross-register settings.

sequence-classification framework[4], which builds on HuggingFace Transformers (Wolf et al., 2020). We use *bert-base-uncased* (Devlin et al., 2019) and *roberta-base* (Liu et al., 2019), both comprising 12 transformer layers with 12 attention heads each. A linear classification head is placed on top of the final hidden state. Models are trained using the Adam optimizer (Kingma and Ba, 2014) and negative log-likelihood loss. We compare performances of BERT and RoBERTa against a majority vote baseline, which always predicts the most frequent class.

Regarding training, we find that two epochs are optimal for fine-tuning on the large USElecDeb corpus across all tasks. In contrast, the smaller size of ArgUNSC benefits from longer training, and we fix the number of epochs between 6 and 8 for all cross-validation runs.

In addition to experimenting with encoder-only models like BERT and RoBERTa, we evaluate a GPT-4 LLM developed by OpenAI (OpenAI, 2023). We prompt GPT-4 under two conditions: zero-shot and few-shot. In the zero-shot setup, the model is given only task instructions without any labeled examples. In the few-shot setup, the prompt is augmented with three labeled examples per class. For instance, in the ACS task, the prompt includes three sentences labeled as arguments and three labeled as non-arguments to guide the model's classification.

4.3 Evaluation Setup

Our experiments are designed to evaluate model performance both within and across two corpora. We consider three main scenarios: (a) fine-tuning and testing on the large-scale USElecDeb corpus (serving as an in-register baseline), (b) fine-tuning and testing on the smaller ArgUNSC corpus, and (c) fine-tuning on USElecDeb and testing on Ar-

[4] https://pypi.org/project/bert-for-sequence-classification/

Evaluation	Label	Majority Vote	BERT			RoBERTa		
		F1	P	R	F1	P	R	F1
IR–USElecDeb$_{\text{sentence}}$	Relation	0.666	0.696	0.824	0.755	0.754	0.876	0.810
	No Relation	0.000	0.785	0.640	0.705	0.852	0.714	0.777
	Avg Weighted	0.333	0.741	0.732	0.730	0.803	0.795	0.794
IR–ArgUNSC$_{\text{sentence}}$	Relation	0.667	0.664 ± 0.023	0.738 ± 0.048	0.697 ± 0.011	0.698 ± 0.022	0.774 ± 0.031	0.733 ± 0.007
	No Relation	0.000	0.706 ± 0.020	0.623 ± 0.061	0.659 ± 0.028	0.747 ± 0.013	0.662 ± 0.048	0.700 ± 0.023
	Avg Weighted	0.333	0.685 ± 0.008	0.680 ± 0.009	0.678 ± 0.011	0.722 ± 0.008	0.718 ± 0.011	0.717 ± 0.012
IR–ArgUNSC$_{\text{component}}$	Relation	0.670	0.642 ± 0.009	0.712 ± 0.040	0.675 ± 0.019	0.708 ± 0.009	0.741 ± 0.062	0.723 ± 0.025
	No Relation	0.000	0.678 ± 0.024	0.603 ± 0.027	0.637 ± 0.011	0.732 ± 0.043	0.693 ± 0.036	0.710 ± 0.006
	Avg Weighted	0.335	0.660 ± 0.014	0.657 ± 0.012	0.656 ± 0.011	0.720 ± 0.019	0.717 ± 0.014	0.717 ± 0.013
CR$_{\text{sentence}}$	Relation	0.667	0.541	0.922	0.682	0.536	0.960	0.688
	No Relation	0.000	0.738	0.219	0.338	0.808	0.169	0.279
	Avg Weighted	0.333	0.640	0.571	0.510	0.672	0.564	0.484

Table 9: F1-score for the majority vote baseline and Precision (P), Recall (R), and F1-scores for BERT and RoBERTa. Task: **Argumentative Relation Identification (ARI)** in in-register and cross-register settings on both sentence and component levels.

gUNSC to evaluate cross-register generalization ability of models. Hereinafter, we adopt the **IR** abbreviation for the *in-register* experiments and **CR** for the *cross-register* ones.

For all experiments involving USElecDeb, we use the official training and testing splits provided by the authors. In the CR setting, we fine-tune the models on the USElecDeb training set and evaluate on the full ArgUNSC corpus. For IR experiments on ArgUNSC, we follow a 5-fold stratified cross-validation protocol and report mean and standard deviation. GPT-4 setups are evaluated using the entire ArgUNSC.

Across all IR and CR settings, and for both encoder-based models and LLMs, each of the four stages in the argument mining pipeline is evaluated using gold standard labels, without propagating errors from one step to the next.

5 Results and Discussion

5.1 In- and Cross-Register Performance

Table 7 presents IR and CR results on the **Argumentative Component Segmentation (ACS)** task. First, we observe that in-register (IR) performance for both ArgUNSC and USElecDeb is moderately high, with RoBERTa approaching an F1 score of 0.8, indicating that argument segmentation (ACS) is a fairly solvable task in both corpora, even with class imbalance. We also note that, with per-class F1 scores of 0.913 and 0.641, our RoBERTa model performs competitively compared to the LSTM predictions reported in Haddadan et al. (2019), which are 0.913 and 0.547, respectively. In the cross-register (CR) setting, RoBERTa achieves an F1 score of 0.783, which—when compared to the strong majority vote baselines in IR and CR—suggests robust generalization, both overall and at the class level. BERT follows a similar pattern, showing solid cross-register performance, although RoBERTa consistently outperforms it across all setups.

Results for the **Argumentative Component Classification (ACC)** task are shown in Table 8. First, we again note a competitive performance of our RoBERTa (0.684) compared to Haddadan et al. (2019)'s LSTM (0.673). Generally, while performance in both IR and CR settings hovers around the 0.7 F1 mark, there is a consistent drop compared to ACS, reflecting the higher complexity of component type classification. Nevertheless, both BERT and RoBERTa generalize remarkably well: in the CR setup, they achieve F1 scores of 0.721 and 0.731 accordingly, surpassing even their IR performance on USElecDeb. This suggests that the conceptual distinction between claims and premises is relatively stable across the two political speech genres.

Argumentative Relation Identification (ARI) results are summarized in Table 9. In the IR set-

Evaluation	Label	Majority Vote	BERT			RoBERTa		
		F1	P	R	F1	P	R	F1
IR–USElecDeb$_{\text{sentence}}$	Support	0.919	0.890	0.970	0.929	0.908	0.954	0.931
	Attack	0.000	0.658	0.326	0.436	0.639	0.456	0.532
	Avg Macro	0.459	0.774	0.648	0.682	0.773	0.705	0.731
IR–ArgUNSC$_{\text{sentence}}$	Support	0.937	0.918 ± 0.008	0.973 ± 0.006	0.944 ± 0.003	0.937 ± 0.014	0.958 ± 0.016	0.947 ± 0.005
	Attack	0.000	0.628 ± 0.039	0.349 ± 0.075	0.444 ± 0.068	0.630 ± 0.047	0.511 ± 0.126	0.551 ± 0.049
	Avg Macro	0.469	0.773 ± 0.022	0.661 ± 0.035	0.694 ± 0.035	0.783 ± 0.021	0.735 ± 0.057	0.749 ± 0.031
IR–ArgUNSC$_{\text{component}}$	Support	0.936	0.925 ± 0.006	0.974 ± 0.007	0.949 ± 0.003	0.943 ± 0.009	0.973 ± 0.010	0.958 ± 0.003
	Attack	0.000	0.673 ± 0.060	0.399 ± 0.055	0.497 ± 0.044	0.741 ± 0.065	0.548 ± 0.075	0.624 ± 0.040
	Avg Macro	0.468	0.799 ± 0.029	0.686 ± 0.026	0.723 ± 0.023	0.842 ± 0.030	0.761 ± 0.034	0.791 ± 0.021
CR$_{\text{sentence}}$	Support	0.937	0.884	0.994	0.936	0.895	0.982	0.937
	Attack	0.000	0.320	0.023	0.043	0.511	0.137	0.216
	Avg Macro	0.469	0.602	0.508	0.489	0.703	0.560	0.576

Table 10: F1-score for the majority vote baseline and Precision (P), Recall (R), and F1-scores for BERT and RoBERTa. Task: **Argumentative Relation Classification (ARC)** in in-register and cross-register settings on both sentence and component levels.

tings, models perform reasonably well, with F1 scores surpassing 0.79 on RoBERTa—comparable to or even exceeding results from ACC, despite ARI typically being considered the more complex task.

In contrast, cross-register generalization (CR) reveals a substantial performance drop: the weighted F1 score decreases to 0.484 for RoBERTa and 0.510 for BERT. Notably, the model barely improves over the majority vote baseline for the Relation class. A likely explanation lies in the structural differences between corpora. In the USElecDeb corpus, argumentative relations are annotated not only between premises and claims but also between claims and between premises. This variation likely introduces noise and confuses the model at inference time.

We also report component-level results for ARI, where we expected a performance gain due to more granular inputs. However, the results remain on par with the sentence-level setting.

Argumentative Relation Classification (ARC) results are reported in Table 10. In the IR settings, both models perform well, but RoBERTa proves to be more competitive. We note that Attack relations remain substantially harder than Support relations, consistently showing F1 scores below 0.65 – even IR.

The CR scenario further highlights this difficulty. While Support generalizes well (0.937 F1 with RoBERTa), Attack F1 drops to 0.216, pulling the macro average down to 0.576 on RoBERTa. These results suggest that although positive argumentative relations transfer reliably across registers, adversarial patterns (e.g., attacks) are less stable.

Component-level results show slightly improved performance compared to sentence-level, with a more pronounced benefit for ARC than ARI. This is likely because fine-grained component boundaries benefit the task of distinguishing relation polarity (support vs. attack) more than the task of relation existence detection.

General findings can be summarized as follows. First, RoBERTa consistently outperforms BERT across all tasks and evaluation settings (with the only notable exception of ARI in cross-register setting), confirming its superior contextual representation capabilities for argumentative language. Second, among the four tasks, ACS emerges as the easiest in the IR setting, likely due to the presence of clear lexical markers. In contrast, ARC proves to be the most challenging, as it demands nuanced modeling of argument polarity. Third, for ARC, moving from sentence- to component-level modeling substantially improves performance, particularly in the ArgUNSC IR setting. RoBERTa achieves near 0.80 F1, underscoring the value of increased granularity in argumentative polarity classification. Finally, regarding generalization, the best transfer is observed for ACC and ACS. ARC ex-

Task	IR	CR	GPT-4 zero	GPT-4 few
ACS	0.819	0.783	0.652	0.767
ACC	0.755	0.733	0.683	0.706
ARI	0.717	0.484	0.594	0.562
ARC	0.749	0.576	0.636	0.639

Table 11: F1 scores (average macro for tasks ACS and ARC and average weighted for tasks ACC and ARI) for GPT-4-prompting methods compared with the IR and CR predictions (RoBERTa).

hibits moderate robustness, suggesting that relation polarity (e.g., support vs. attack) transfers more reliably than the identification of whether a relation exists at all. ARI remains the most difficult to generalize, potentially due to cross-register differences in density, directionality, and linking strategies in underlying argument structure graphs.

5.2 Comparison to GPT-4 models

Table 11 presents the results of zero-shot and few-shot prompting using the GPT-4 model across the four core argument mining tasks on ArgUNSC dataset. Overall, GPT-4 underperforms compared to fine-tuned RoBERTa models in all IR and half of the CR scenarios. The gap is particularly pronounced in tasks ACS (0.819 IR and 0.783 CR vs. 0.767 few-shot) and ACC (0.755 IR and 0.733 CR vs 0.706 few-shot). On ARI and ARC, zero- and few-shot prompting outperforms the CR setup, but it still falls short of IR-fine-tuned RoBERTa on these tasks with.

This may be because fine-tuned BERT-based models are directly adapted to the domain, register and context intricacies of the dataset, while prompting alone often fails to capture such subtleties—especially for complex discourse tasks like argumentative relation detection in political speech. Fine-tuning large open-weight models such as LLaMA (Touvron et al., 2023) or Mistral (Jiang et al., 2023) could address this gap.

6 Conclusion

Our work presents a comprehensive study of cross-register generalization in argument mining within political discourse. We introduce ArgUNSC, a new manually annotated corpus of UN Security Council speeches, and benchmark four core AM tasks.

We acknowledge several limitations. The study is restricted to the English language and two political registers. Further, our sentence-level setup simplifies structures in multi-component sentences,—future work may explore more fine-grained approaches, such as token-level prediction.

In the future, we plan to conduct a qualitative error analysis to identify which register-specific differences contribute to model failures in ARI and ARC.

Beyond its value for argument mining pipelines, ArgUNSC also opens new avenues for political science research, such as analyzing how nations justify their own or foreign policies and rhetorically align with allies or opponents.

Reproducibility

The new ArgUNSC dataset, annotation guidelines and Python scripts can be found at: https://github.com/mpoiaganova/political-argument-mining

Acknowledgements

The work was supported by the Deutsche Forschungsgemeinschaft (DFG), project (448421482) "Trajectories of Conflict: The Dynamics of Argumentation in the UN Security Council“.

We thank Costanza Rasi, Dietmar Benndorf, and Karolina Zaczynska for their support with data annotation and corpus preparation.

References

Yamen Ajjour, Wei-Fan Chen, Johannes Kiesel, Henning Wachsmuth, and Benno Stein. 2017. Unit segmentation of argumentative texts. In *4th Workshop on Argumentation Mining*, pages 118–128.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational linguistics*, 34(4):555–596.

Nico Blokker, Erenay Dayanik, Gabriella Lapesa, and Sebastian Padó. 2020. Swimming with the tide? positional claim detection across political text types. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 24–34, Online. Association for Computational Linguistics.

Jérémie Cabessa, Hugo Hernault, and Umer Mushtaq. 2025. Argument mining with fine-tuned large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6624–6635, Abu Dhabi, UAE. Association for Computational Linguistics.

Tuhin Chakrabarty, Christopher Hidey, Smaranda Muresan, Kathy McKeown, and Alyssa Hwang. 2019. AMPERSAND: Argument mining for PERSuAsive oNline discussions. 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 2933–2943, Hong Kong, China. Association for Computational Linguistics.

Guizhen Chen, Liying Cheng, Anh Tuan Luu, and Lidong Bing. 2024. Exploring the potential of large language models in computational argumentation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2309–2330, Bangkok, Thailand. Association for Computational Linguistics.

Liying Cheng, Lidong Bing, Ruidan He, Qian Yu, Yan Zhang, and Luo Si. 2022. IAM: A comprehensive and large-scale dataset for integrated argument mining tasks. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2277–2287, Dublin, Ireland. Association for Computational Linguistics.

Johannes Daxenberger, Steffen Eger, Ivan Habernal, Christian Stab, and Iryna Gurevych. 2017. What is the essence of a claim? cross-domain claim identification. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2055–2066, Copenhagen, Denmark. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Lucile Favero, Juan Antonio Pérez-Ortiz, Tanja Käser, and Nuria Oliver. 2025. Leveraging small llms for argument mining in education: Argument component identification, classification, and assessment. *arXiv preprint arXiv:2502.14389*.

Marc Feger and Stefan Dietze. 2024. TACO – Twitter arguments from COnversations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 15522–15529, Torino, Italia. ELRA and ICCL.

Debela Gemechu, Ramon Ruiz-Dolz, and Chris Reed. 2024. Aries: A general benchmark for argument relation identification. In *11th Workshop on Argument Mining, ArgMining 2024*, pages 1–14. Association for Computational Linguistics (ACL).

Pierpaolo Goffredo, Shohreh Haddadan, Vorakit Vorakitphan, Elena Cabrio, and Serena Villata. 2022. Fallacious argument classification in political debates. In *Thirty-First International Joint Conference on Artificial Intelligence {IJCAI-22}*, pages 4143–4149. International Joint Conferences on Artificial Intelligence Organization.

Deniz Gorur, Antonio Rago, and Francesca Toni. 2025. Can large language models perform relation-based argument mining? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8518–8534, Abu Dhabi, UAE. Association for Computational Linguistics.

Ivan Habernal, Daniel Faber, Nicola Recchia, Sebastian Bretthauer, Iryna Gurevych, Indra Spiecker genannt Döhmann, and Christoph Burchard. 2024. Mining legal arguments in court decisions. *Artificial Intelligence and Law*, 32(3):1–38.

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.

Dieuwke Hupkes, Mario Giulianelli, Verna Dankers, Mikel Artetxe, Yanai Elazar, Tiago Pimentel, Christos Christodoulopoulos, Karim Lasri, Naomi Saphra, Arabella Sinclair, et al. 2023. A taxonomy and review of generalization research in nlp. *Nature Machine Intelligence*, 5(10):1161–1174.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart De Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th international conference on computational linguistics: system demonstrations*, pages 5–9.

Marco Lippi and Paolo Torroni. 2016a. Argument mining from speech: Detecting claims in political debates. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30.

Marco Lippi and Paolo Torroni. 2016b. Margot: A web server for argumentation mining. *Expert Systems with Applications*, 65:292–303.

Boyang Liu, Viktor Schlegel, Riza Batista-Navarro, and Sophia Ananiadou. 2023. Argument mining as a multi-hop generative machine reading comprehension task. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10846–10858, Singapore. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. 2018. Never retreat, never retract: Argumentation analysis for political speeches. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Rafael Mestre, Razvan Milicin, Stuart E. Middleton, Matt Ryan, Jiatong Zhu, and Timothy J. Norman. 2021. M-arg: Multimodal argument mining dataset for political debates with audio and transcripts. In *Proceedings of the 8th Workshop on Argument Mining*, pages 78–88, Punta Cana, Dominican Republic. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report.

Ramon Ruiz-Dolz and John Lawrence. 2023. Detecting argumentative fallacies in the wild: Problems and limitations of large language models. In *Proceedings of the 10th Workshop on Argument Mining*. Association for Computational Linguistics.

Robin Schaefer, René Knaebel, and Manfred Stede. 2022. On selecting training corpora for cross-domain claim detection. In *Proceedings of the 9th workshop on argument mining*, pages 181–186.

Mirco Schönfeld, Steffen Eckhard, Ronny Patz, and Hilde Van Meegdenburg. 2019. The un security council debates 1995-2017. *arXiv preprint arXiv:1906.10969*.

Noam Slonim, Yonatan Bilu, Carlos Alzate, Roy Bar-Haim, Ben Bogin, Francesca Bonin, Leshem Choshen, Edo Cohen-Karlik, Lena Dankin, Lilach Edelstein, et al. 2021. An autonomous debating system. *Nature*, 591(7850):379–384.

Ekaterina Sviridova, Anar Yeginbergen, Ainara Estarrona, Elena Cabrio, Serena Villata, and Rodrigo Agerri. 2024. CasiMedicos-arg: A medical question answering dataset annotated with explanatory argumentative structures. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18463–18475, Miami, Florida, USA. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Jacky Visser, Barbara Konat, Rory Duthie, Marcin Koszowy, Katarzyna Budzynska, and Chris Reed. 2020. Argumentation in the 2016 us presidential elections: annotated corpora of television debates and social media reaction. *Language Resources and Evaluation*, 54(1):123–154.

Thiemo Wambsganss, Tobias Kueng, Matthias Soellner, and Jan Marco Leimeister. 2021. Arguetutor: An adaptive dialog-based learning system for argumentation skills. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–13.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Karolina Zaczynska, Peter Bourgonje, and Manfred Stede. 2024. How diplomats dispute: The un security council conflict corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8173–8183.

A ArgUNSC Descriptive Statistics

Year	Speeches
2014	93
2015	27
2016	11
2017	7
2018	6

Table 12: Number of speeches per year

Country	**Speeches**
Russia	25
Ukraine	16
United States	15
United Kingdom	11
France	11
China	11
Lithuania	8
Australia	7
Rwanda	6
The Republic of Korea	6
Luxembourg	5
Argentina	4
Chile	4
Nigeria	3
Jordan	2
Sweden	1
Ethiopia	1
Angola	1
Belgium	1
New Zealand	1
Venezuela	1
Spain	1
Chad	1
Indonesia	1
UNSC Briefing	1

Table 13: Number of speeches per country