Argument Structure in Online Debates: Proposal

Aratrik Paul
UC Berkeley
School of Information

Minkush Jain
UC Berkeley
School of Information

Vikramsingh Rathod
UC Berkeley
School of Information

Abstract

Online debates encode claims, counter-claims, and supporting evidence that shape public discourse, yet most NLP work emphasizes stance or sentiment rather than the *structure* of arguments. We propose to mine argument components (claims, premises) and relations (support, attack) in debate corpora, then compare how pro versus con sides structure their reasoning. Our contributions will be (i) a reproducible pipeline for component and relation detection, (ii) an analysis of rhetorical strategy differences across sides, and (iii) an error taxonomy useful for future argument mining research.

1 Introduction & Motivation

Debate platforms and social media threads contain rich argumentative structure that influences persuasion and polarization. While stance detection has matured, there is comparatively less empirical work that models how *arguments are built*: which claims are advanced, what evidence appears as premises, and how support/attack relations differ across opposing sides. We ask:

> **Research Question 1:** How do the structures of arguments differ between pro and con positions within the same debate topic?
>
> **Research Question 2:** Which component/relationship patterns (e.g - frequency of attacks, depth of support chains) characterize each side's rhetorical strategy?

By answering these questions, we aim to surface linguistic mechanisms underlying persuasion and disagreement in online debates. Beyond predictive accuracy, we emphasize interpretable analyses that connect NLP outputs to rhetorical behavior.

2 Related Work & Gap

Argument mining systems detect components (claims, premises) and relations between them. Transition-based and graph-centric approaches incrementally construct argument graphs without exhaustive pairwise linking, improving efficiency and robustness across structures (Bao et al., 2021). Applications to noisy, short-form discourse (e.g., Twitter) demonstrate feasibility but often frame tasks as coarse classification (e.g., reasons present vs. absent) rather than full structure extraction (Bhatti et al., 2021). Foundational corpora of persuasive essays provide strong supervision for component and relation labels (Stab and Gurevych, 2017). Quality assessment research connects argument features to rhetorical dimensions such as clarity and sufficiency (Wachsmuth et al., 2018). Surveys outline major challenges, including fragmented arguments in social media and domain transfer (Lawrence and Reed, 2019).

Gap: Few studies explicitly compare *structural* differences across opposing sides within the same debate (e.g., relative use of attacks vs. supports, evidence breadth, or graph depth). Our work addresses this gap through side-aware structure mining and quantitative rhetorical analysis.

Datasets

Primary dataset - Persuasive Essays (Stab and Gurevych, 2017):
Essays annotated with claims, premises, and support/attack relations; provides high-quality span and relation labels for training and evaluation.

Secondary dataset - IBM Debater AM (Bao et al., 2021):
Argument mining corpus with component and

relation annotations across topics, enabling cross-domain robustness checks.

Optional extension - ChangeMyView (CMV) Reddit:
Arguments with stance metadata; if used, we will adapt labels to our schema and treat CMV as an out-of-domain test set.

2.1 Preprocessing & EDA

We will compute label distributions, span lengths, class imbalance, and relation graph density per topic and per side (when stance is available). We will either adopt official splits or create stratified train/dev/test splits by topic to assess generalization.

3 Methods

Baselines

- **Component tagging:** BiLSTM–CRF sequence tagger for claim/premise spans.
- **Relation classification:** Logistic regression or linear SVM on pair features (distance, lexical overlap) as a simple baseline.

Transformer models

- **Token/Span classification:** Fine-tune BERT/Longformer for component detection.
- **Relation modeling:** Cross-encoder over sentence/segment pairs to classify SUPPORT/ATTACK/NONE.

Graph construction Following Bao et al. (2021), we will explore a lightweight transition-based parser to build argument graphs, comparing against a pipeline that first predicts components then links relations.

Side-aware analyses When stance labels are available (e.g., by topic split or corpus metadata), we will partition instances into PRO and CON and compute:

- component mix (claims vs. premises) and evidence breadth,
- relation mix (SUPPORT vs. ATTACK) and graph depth,
- lexical and discourse markers (e.g., causal connectives, citations).

4 Evaluation

Structure metrics

- **Components:** span-level precision, recall, F1 (exact and partial match).
- **Relations:** macro-F1 over SUPPORT/ATTACK/NONE; edge accuracy.

Generalization & robustness Topic-held-out evaluation within a corpus; optional cross-domain test (train on essays, test on IBM Debater AM).

Rhetorical findings We will report statistically tested differences between sides (e.g., proportion of attack edges, average support-chain depth, evidence density per claim). A small human sanity-check (n=50 instances) will validate error categories and interpretability of patterns.

5 Expected Contributions

1. A reproducible pipeline for argument component and relation extraction in debates.
2. The first (to our knowledge) side-aware comparison of structural argument patterns within topics across corpora.
3. An error taxonomy and analysis artifacts (tables/plots) that future work can reuse.

6 Ethics & Limitations

We will use public, licensed datasets and follow data-use policies. Debate texts may contain sensitive content; we will avoid releasing any personally identifying information and share only aggregate analyses and trained models where licenses permit. Limitations include domain mismatch (essays vs. online debates), annotation subjectivity for relations, and potential stance skew by topic. Our emphasis on transparent error analysis and side-aware reporting mitigates over-interpretation.

Team Responsibilities

Aratrik Paul (literature review & modeling co-lead):

- Lead related work synthesis and contribute to proposal + final report writing.

- Implement one baseline argument mining model (e.g., BiLSTM) and run evaluation on small datasets.
- Coordinate weekly team check-ins and manage Overleaf draft.

Minkush Jain (data & analysis co-lead):

- Handle dataset acquisition, cleaning, preprocessing, and exploratory data analysis (EDA).
- Implement dataset statistics + visualization pipeline (corpus stats, label distributions).
- Take meeting notes and track internal deadlines.

Vikramsingh Rathod (experiments & evaluation co-lead):

- Implement transformer-based model (e.g., RoBERTa fine-tuning) and ablation experiments.
- Develop evaluation scripts (accuracy, F1, argument relation metrics) and error analysis.
- Compile slides/demos for class presentations and final submission.

Milestones and Timeline

- **October 6:** Finalize dataset selection (Persuasive Essays, IBM Debater AM, CMV extension) and confirm preprocessing schema.
- **October 13:** Complete dataset cleaning, preprocessing, and exploratory data analysis (EDA).
- **October 20:** Implement baseline models (BiLSTM, logistic regression) and generate initial evaluation metrics.
- **October 27:** Literature review synthesis complete; submit first draft of related work section.
- **November 3:** Fine-tune transformer model (RoBERTa/BERT) on primary dataset; record results.
- **November 10:** Conduct cross-domain experiments using IBM Debater AM; prepare error analysis notes.
- **November 17:** Run ablation studies (features, architectures); finalize evaluation scripts.
- **November 24:** Integrate results into Overleaf draft; circulate for team-wide editing and feedback.
- **December 1:** Prepare final figures, tables, and polished draft; rehearse presentation.
- **December 4:** Submit final paper and deliver class presentation.

References

Jianzhu Bao, Chuang Fan, Jipeng Wu, Yixue Dang, Jiachen Du, and Ruifeng Xu. 2021. A neural transition-based model for argumentation mining. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6354–6364. Association for Computational Linguistics.

Muhammad Mahad Afzal Bhatti, Ahsan Suheer Ahmad, and Joonsuk Park. 2021. Argument mining on twitter: A case study on the planned parenthood debate. In *Proceedings of the 8th Workshop on Argument Mining*, pages 1–11. Association for Computational Linguistics.

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

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Henning Wachsmuth, Khalid Al-Khatib, and Benno Stein. 2018. Quality assessment for argumentation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1680–1691. Association for Computational Linguistics.