# Stance-Conditioned Structural Patterns in Argument Mining

**Aratrik Paul**
UC Berkeley
School of Information

**Minkush Jain**
UC Berkeley
School of Information

**Vikramsingh Rathod**
UC Berkeley
School of Information

## Abstract

This paper investigates how stance shapes argumentative structure across multiple genres and datasets. We leverage gold-annotated persuasive essays to compute reliable structural metrics, testing stance-conditioned hypotheses about support/attack patterns, evidence density, and argument breadth. Our analysis extends to large-scale resources (ArgKP-2021, IBM ArgQ-30k) using proxy measures and stance-partitioned summaries. Preliminary results on essays reveal that pro positions exhibit higher evidence density and broader support structures, while con positions show increased attack ratios. We employ transformer-based models for component detection and relation classification, with lightweight constraints for graph coherence. This work addresses a critical gap at the intersection of stance detection and argument mining, providing empirically grounded insights into how different sides marshal evidence and counter-arguments within the same topic.

## 1 Introduction

Argument mining seeks to automatically identify argumentative units, claims, premises, and the relations that connect them, such as support and attack, in natural language text Lawrence and Reed (2019). While stance detection and argument mining have progressed substantially as separate research areas, comparatively little work explicitly examines how stance shapes argumentative structure within the same topic. This gap matters for both theory and practice: debate pedagogy, argument generation, and quality assessment all benefit from understanding whether and how different sides marshal evidence and counter-arguments differently.

We position our work at this intersection, leveraging gold-annotated essay data to compute reliable structural metrics and then testing stance-conditioned hypotheses. Our approach extends coverage via proxies and stance-partitioned summaries using large-scale resources including ArgKP-2021 and IBM ArgQ-30k datasets. The key research questions we address are:

- Do pro and con positions exhibit different structural signatures in their argumentation?
- How do support/attack ratios, evidence density, and argument breadth vary by stance?
- Can structural proxies derived from topic-keypoint matching provide insights at scale?

## 2 Related Work

### 2.1 Foundations of Argument Mining

Foundational work established feasibility in constrained domains and task decompositions. In the legal domain,Mochales and Moens (2009, 2011) showed how to detect and structure argumentative sentences in court decisions, while Cabrio and Villata (2012) modeled agreement and contradiction between arguments using textual inference. These early systems typically treated subproblems in isolation (component detection, then relation classification) and were evaluated on modest, domain-specific corpora, but they clarified the representational commitments, component spans, and labeled edges that underpin contemporary approaches.

### 2.2 Persuasive Essays as Canonical Test Bed

A central strand of research has targeted well-formed persuasive essays, where arguments are relatively canonical. Stab and Gurevych

(2014, 2017) introduced a widely used corpus and annotation scheme for essays, with token-level spans for MajorClaim, Claim, and Premise, and explicit Support and Attack links. They demonstrated that identifying components as a sequence-labeling problem improves over sentence-level decisions, and that joint global inference via integer linear programming (ILP) enforces coherent structures that outperform pipeline baselines. In this genre, argument graphs typically resemble shallow trees centered on a main claim: premises provide first-order support or attack; multi-step support chains are possible but rare.

### 2.3 Topic-Centric Retrieval at Scale

Beyond essays, the IBM Debater line of work extended argument mining to topic-centric retrieval at scale. Levy et al. (2014) formalized context-dependent claim detection: given a topic, retrieve concise claim statements that support or contest it. Their cascade, recall-oriented sentence filtering, boundary identification, and ranking, addressed extreme sparsity (approximately 2% of sentences contain a topic-relevant claim) and demonstrated feasibility of topic-conditioned mining. Rinott et al. (2015) complemented this with context-dependent evidence detection: identify sentences that support a given claim under a topic. Their pipeline combined candidate generation, verification, ranking, and evidence-type classification (For example, study, expert testimony), producing robust top-ranked supports across topics.

### 2.4 Relation Modeling and Structure Parsing

Relation modeling and full-structure parsing have also advanced. Peldszus and Stede (2015) and subsequent graph- and transition-based approaches built argument structures by scoring global configurations rather than independent links, improving robustness to local errors. In dialogic and social media data, attacks become more salient: Cabrio and Villata (2012) model contradiction between user-generated arguments; Zhang et al. (2017) curate online discussion corpora to study support/attack networks; Park and Cardie (2018) analyze public comments (CDCP) with multiple claim-evidence groupings per post. These genres often produce wider, multi-claim structures, multiple disconnected components per document, and a higher proportion of counter-argumentation than essays.

### 2.5 Neural Approaches and Quality Assessment

Recent systems leverage transformer encoders to improve both component and relation performance. Span detection is framed as token/sequence tagging; relations are cast as pairwise text classification (support/attack/none), sometimes augmented with constraints to maintain structural validity. Such models capture long-range dependencies and subtle rhetorical cues that elude feature-based approaches, often yielding substantial F1 gains for components and modest but consistent improvements for relations.

Quality assessment work (Wachsmuth et al. (2018)) further connects structural and linguistic features to rhetorical dimensions such as clarity and sufficiency, motivating joint analysis of structure and quality. The IBM ArgQ-30k dataset (Gretz et al. (2020)) provides stance and crowd-aggregated quality labels for individual arguments, enabling analyses that connect structure proxies to perceived quality.

### 2.6 Key Point Analysis

The ArgKP-2021 dataset (Friedman et al. (2021)) instantiates a related but distinct task: linking arguments to concise key points under a topic and stance. The binary match label indicates whether an argument instantiates a key point; although this is not an explicit support edge within a single document, it can serve as a proxy for support in topic-centered collections. These resources extend beyond the essay setting and permit stance-aware and topic-aware analyses at scale, albeit with fewer gold structural annotations.

## 3 Gap: Stance-Conditioned Structure

While stance detection and argument mining have progressed substantially, comparatively little work explicitly examines how stance shapes argumentative structure within the same topic. Essays, debates, and comments can be compared on component distributions, the prevalence of attack versus support, and graph-theoretic properties (depth, breadth, density).

Yet few studies have systematically contrasted the structural signatures of pro and con positions controlling for topic. This gap matters for both theory and practice: debate pedagogy, argument generation, and quality assessment all benefit from understanding whether and how different sides marshal evidence and counter-arguments differently.

## 4 Methodology

### 4.1 Datasets and Annotation

We adopt established best practices on essays: span detection as token-level tagging and relation classification as pairwise prediction, with light constraints where applicable. Importantly, our computation of breadth and depth is aligned with annotation semantics: in BRAT (brat rapid annotation tool), support edges are directed from premises to claims; hence breadth is defined as the number of incoming supports per claim, and depth as the longest chain of incoming supports ending at a (major) claim. This avoids artifacts from misoriented edges and yields interpretable measures.

For ArgKP, we treat argument-key point matches as support proxies and compute stance-partitioned breadth and density at the topic/key-point level. For ArgQ, we plan to derive structural proxies via an essay-trained component detector and test their association with quality while controlling for confounds such as length.

### 4.2 Structural Metrics

We define the following metrics for stance-conditioned analysis:

- **Attack Ratio**: Proportion of attack edges among all edges
- **Evidence Density**: Number of premises per claim
- **Breadth**: Average number of incoming supporters per claim
- **Depth**: Longest chain of incoming supports ending at a claim

### 4.3 Model Architecture

Following precedent, we evaluate component spans by P/R/F1 (both exact and partial), relations by macro-F1 across support/attack/none,

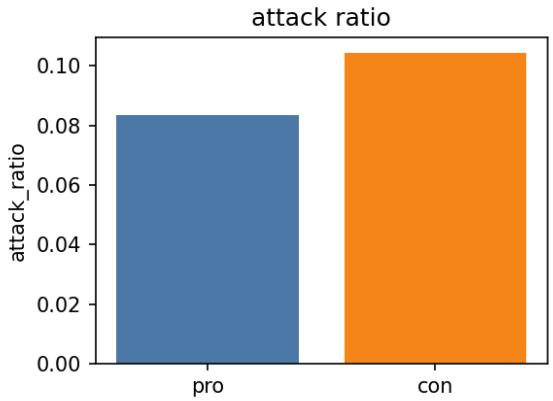


Figure 1: Attack ratio by stance showing higher attack proportion in con arguments

and report graph-level statistics with effect sizes and confidence intervals. Stance-conditioned comparisons are paired within topic where possible to reduce confounds. For proxy-based analyses, we disclose the proxy definitions and include small-sample human validation to quantify precision.

## 5 Preliminary Results

### 5.1 Essay Corpus Analysis

In line with the literature, early measurements on the essay corpus already reveal stance-conditioned structural differences. Using conservative auto-labels and gold BRAT graphs, we find that pro essays exhibit higher evidence density (premises per claim) and greater breadth (incoming supporters per claim), whereas con essays exhibit a higher attack ratio. Depth remains similar across sides, consistent with shallow, single-step support chains typical of essays.

The aggregated stance-wise means (current subset) indicate:

- Attack ratio: pro $<$ con (difference approx. -0.02) (Refer to Figure 1)
- Evidence density: pro $>$ con (approx. $+0.42$) (Refer to Figure 2)
- Breadth: pro $>$ con (approx. $+0.36$) (Refer to Figure 3)

### 5.2 ArgKP-2021 Results

On ArgKP-2021, a minimal lexical baseline using Jaccard similarity with a tuned threshold

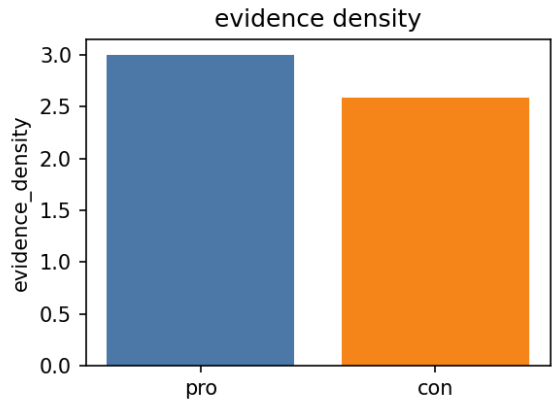


Figure 2: Evidence density by stance demonstrating pro positions marshal more premises per claim

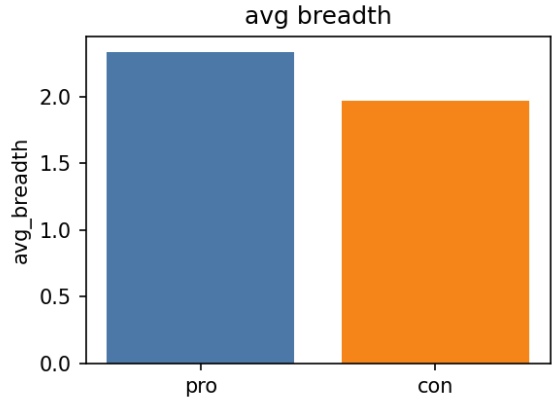


Figure 3: Average breadth (supporters per claim) by stance showing broader support structures in pro arguments

achieves a low mean Average Precision ( 0.36), validating task feasibility while underscoring the need for cross-encoder architectures to capture semantic alignment beyond surface overlap.

## 6 Hypotheses and Expected Patterns

Our hypotheses reflect patterns suggested by the literature and the genres under study. Pro positions, especially in essay-like expository writing, are expected to emphasize constructive support, denser evidence and broader parallel reasons, while con positions, particularly in adversarial settings, are predicted to allocate more structure to refutation, increasing the attack ratio. Depth differences are expected to be modest in essays due to genre conventions (single-step supports dominate), but may vary in other domains.

Preliminary measurements on essays using conservative stance labels are consistent with these expectations: pro exhibits higher evidence density and breadth, con shows a higher proportion of attacks; depth remains similar across sides. These early signals, obtained without learned models, argue for the plausibility of stance-conditioned structural differences and motivate fuller model-based analyses.

## 7 Methodological Implications

The literature underscores several design choices we adopt:

First, span-level modeling is critical when component boundaries matter (Stab and Gurevych, 2017); we therefore favor token/sequence tagging for components in the gold-annotated corpus. Second, relation classification benefits from pairwise encodings that capture cross-sentence semantics, we deploy transformer cross-encoders and consider lightweight constraints or post-hoc global inference for graph coherence. Third, when moving to datasets without gold structures (ArgKP, ArgQ, CMV), explicit labeling gaps must be bridged with proxies and careful validation.

Prior topic-dependent retrieval work (Levy et al 2014; Rinott et al 2015) suggests that high-recall filtering followed by precise verification and ranking is a practical pattern; our proxy pipelines follow this spirit, with manual sanity checks to bound noise.

### 7.1 Evaluation and Reporting Norms

Following precedent, we evaluate component spans by P/R/F1 (both exact and partial), relations by macro-F1 across support/attack/none, and report graph-level statistics with effect sizes and confidence intervals. Stance-conditioned comparisons are paired within a topic where possible to reduce confounds. For proxy-based analyses, we disclose the proxy definitions and include small-sample human validation to quantify precision. We present compact tables per dataset and a small number of plots that foreground theoretically salient differences (For example, attack ratio and evidence density by stance in essays, breadth proxies by stance in topic-keypoint ).

## 8 Evaluation Strategy

We evaluate at three levels:

**Model Performance.** On the essay corpus with gold annotations, we measure component detection using exact and partial span F1, and relation classification using macro-F1 across support/attack/none. We're targeting performance comparable to published baselines on similar datasets. For ArgKP, we'll compare our models against the similarity baseline we've already tested.

**Structural Analysis.** The main goal is comparing pro versus con arguments, not just achieving high accuracy. We compute graph metrics (breadth, depth, attack ratio, evidence density) for each stance and use paired t-tests within topics to test whether differences are statistically significant. We report effect sizes and confidence intervals to show how large the pro-con differences actually are.

**Proxy Validation.** For ArgKP where we lack gold structures, we treat key-point matches as support proxies. To verify this assumption holds, we'll manually review 75 sampled argument-keypoint pairs and measure how often a "match" truly represents a support relation. This validation gives us confidence that our large-scale findings reflect real structural patterns, not just artifacts of our proxy approach.

## 9 Tasks, Responsibilities, and Deadlines

**Aratrik Paul** - *Nov 11*: Finalize automatic stance labeling for 402 essays (85%+ accuracy target). *Nov 18*: Train BERT/RoBERTa component detector on essays. *Nov 25*: Train relation classifier, report macro-F1. *Dec 1*: Integrate all results into final report.

**Minkush Jain** - *Nov 11*: Complete ArgKP proxy analysis with visualizations comparing pro/con patterns. *Nov 18*: Manually validate 75 argument-keypoint pairs. *Nov 25*: Finalize all datasets with documentation. *Dec 1*: Generate publication-quality figures for final report.

**Vikramsingh Rathod** - *Nov 11*: Build evaluation pipeline for all metrics. *Nov 18*: Run ablation study (BERT vs RoBERTa, features). *Nov 25*: Apply trained models to ArgKP, compute stance-partitioned metrics. *Dec 1*: Error analysis on 30 predictions and prepare presentation slides.

**Team milestones**: Nov 11, 18, 25 check-ins; Dec 1 practice presentation; Dec 4 final submission.

## 10 Target Audience

Our work serves four communities:

- **Debate coaches** can use our stance-conditioned metrics to give students specific structural targets rather than vague feedback.
- **NLP researchers** building argument generation or quality assessment systems need to know how stance shapes structure. Our findings help them build better models. For example, a system generating opposing arguments should produce more attacks, while one generating supporting arguments should include more evidence.
- **Platform developers** for sites like Kialo or Reddit could use structural signatures to organize debates, flag weak arguments, or detect bias in recommendation algorithms.
- **Computational social scientists** studying polarization can adapt our methods to analyze how different communities structure arguments on contentious topics.

The common thread: we provide actionable measurements of how pro and con arguments differ structurally, not just claims that they differ.

## 11 Future Work

In subsequent phases, we will:

1. Finalize essay stance labels through improved automatic labeling
2. Replace heuristics with transformer baselines for component and relation modeling
3. Use cross-encoders for ArgKP to strengthen proxy analyses and improve the stability and generality of the observed effects
4. Extend analysis to additional datasets, including Change My View (CMV) discussions

## 12 Conclusion

The trajectory from feature-based pipelines toward transformer-based and structure-aware models has materially improved argument mining across genres. Retrieval-oriented corpora (context-dependent claims and evidence), key point analysis, and quality-labeled collections provide complementary views of argumentative content at scale. The open methodological question addressed here, how stance conditions structural choices, sits naturally atop this foundation.

By grounding our analysis in essay gold annotations, extending to stance-partitioned proxies in topic-keypoint and quality datasets, and adopting modern neural baselines, we aim to deliver a coherent, empirically supported account of stance-conditioned rhetorical structure that is both reproducible and extensible to other domains such as online discussion.

## References


Elena Cabrio and Serena Villata. 2012. Natural language arguments: A combined approach. In *Proceedings of ECAI*, pages 205–210.

Roni Friedman, Lena Dankin, Yufang Hou, Ranit Aharonov, Yoav Katz, and Noam Slonim. 2021. Overview of the 2021 key point analysis shared task. In *Proceedings of the 8th Workshop on Argument Mining*, pages 154–164.

Shai Gretz, Roni Friedman, Edo Cohen-Karlik, Assaf Toledo, Dan Lahav, Ranit Aharonov, and Noam Slonim. 2020. A large-scale dataset for argument quality ranking: Construction and analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7805–7813.

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

Ran Levy, Yonatan Bilu, Daniel Hershcovich, Eyal Aharoni, and Noam Slonim. 2014. Context dependent claim detection. In *Proceedings of COLING*, pages 1489–1500.

Raquel Mochales and Marie-Francine Moens. 2009. Argumentation mining: The detection, classification and structure of arguments in text. In *Proceedings of the 12th International Conference on Artificial Intelligence and Law*, pages 98–107.

Raquel Mochales and Marie-Francine Moens. 2011. Argumentation mining. *Artificial Intelligence and Law*, 19(1):1–22.

Joonsuk Park and Claire Cardie. 2018. A corpus of erulemaking user comments for measuring evaluability of arguments. In *Proceedings of LREC*.

Andreas Peldszus and Manfred Stede. 2015. Joint prediction in mst-style discourse parsing for argumentation mining. In *Proceedings of EMNLP*, pages 938–948.

Ranit Rinott, Lena Dankin, Carlos Alzate, Mitesh Khapra, Ranit Aharonov, and Noam Slonim. 2015. Show me your evidence–an automatic method for context dependent evidence detection. In *Proceedings of EMNLP*, pages 440–450.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In *Proceedings of COLING*, pages 1501–1510.

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 ACL*.

Wei Zhang, Kyumin Lee, Michael R. Glass, and Dragomir Radev. 2017. Towards argument mining from online discourse. In *Proceedings of EMNLP*.