Towards Complex Debate Understanding: Predicting Claim Impact Scores Through the Modelling of Claim Interactions

Maxime Brouat[1], **Mihai Surdeanu**[2], **Srdjan Vesic**[1], **Eduardo Blanco**[2]
[1]CRIL, CNRS, Univ. Artois, Lens, France
{brouat, vesic}@cril.fr

[2]University of Arizona, Tucson, AZ, USA
{msurdeanu, eduardoblanco}@arizona.edu

Abstract

Structured debates can be naturally modeled as argument graphs, with claims connected by support and attack relations, a representation formalised in Computational Argumentation Theory. In this paper, we propose a novel neural architecture that jointly models both the textual content of claims and their relational structure. Claims are encoded using contextualised embeddings and compressed through a feedforward compression layer. Then, a graph attention network explicitly captures attack/support interactions. Trained on real-world debates from the Kialo platform, our model predicts the distribution of user-assigned impact votes for each claim. It achieves a mean absolute error (MAE) of 0.068, significantly outperforming both text-only and structure-only baselines. Further experiments show strong out-of-domain generalisation across thematic clusters, as well as suggestive correlations between the model's attention patterns and human voting behaviour. An analysis of linguistic and graph-based features suggests that the model relies on latent argumentative patterns as well as the text. Our findings also shed light on language differences between strong and weak claims, as determined by humans as well as by our best model. All resources and code are openly available at this repository.

Keywords: Discourse Annotation, Representation and Processing, Language Representation Models, Opinion Mining/Sentiment Analysis, Statistical and Machine Learning Methods

1. Introduction

The task of natural language inference (NLI), i.e., deciding whether a hypothesis can be inferred from a premise, is a central natural language problem with applications in medicine (Romanov and Shivade, 2018), law (Kwak et al., 2022; Koreeda and Manning, 2021), and science (Sadat and Caragea, 2022). However, real-world debates are more complex, involving many arguments that interact through *support* (pro) and *attack* (con) relations. While argument graphs have been extensively studied from a theoretical perspective in Computational Argumentation Theory (CAT) (Baroni et al., 2018), their empirical investigation in natural language processing remains limited. NLP approaches to argumentation have often emphasised argument mining, i.e., automatically extracting argument components and pairwise stances from raw text, whereas CAT has concentrated on the formal semantics of non-textual argument graphs, where arguments are treated as abstract units and only their relations are modelled. Yet these two perspectives rarely meet in practice. Our work addresses this gap by modelling user-assigned relevance scores for each claim, referred to as impact distributions, in structured debate graphs. We assume such graphs are already available and focus on their semantics, going beyond pairwise stance classification to reason over the full set of support and attack relations.

While the underlying neural components are established techniques, their integration into a unified architecture for argumentation graphs is, to our knowledge, novel. Combining contextualised sentence embeddings with structure-aware message passing enables us to capture both textual semantics and relational dynamics. Together with the release of debate graphs and their associated impact annotations, this positions our work at the intersection of CAT and empirical NLP.

Recent work (Savigny and Yun, 2025) advances multi-task argument mining with LLMs, focusing on pairwise relations rather than full argumentative structures, thus motivating the need for structure-aware models. In this paper, we present a method to model structured debates by combining textual and relational information. Our approach (a) learns contextualised representations of individual argument texts with Sentence-BERT (SBERT) (Reimers and Gurevych, 2019), and (b) captures attack/support interactions with a graph attention network (GAT) (Veličković et al., 2018). Both components are trained jointly on debates from Kialo, a collaborative argumentation platform.[1] In accordance with the Kialo license, we release only the URLs of the debates used, together with scripts for their conversion into graph format.

The main contributions of this paper are:

- We are among the first to formulate the prob-

[1]https://www.kialo-edu.com

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 8291–8302
11–16 May 2026.

lem of assigning scores to all claims in user-generated debates (181 debates, 459 claims and 26 tokens per claim on average). Specifically, we propose a neural architecture that couples semantic representations of the claims and GAT to account for debate structure, including support and attack relations.
- We conduct comprehensive experiments demonstrating that our approach (a) outperforms alternatives grounded on either the text in the claims or debate structure, and (b) is robust when evaluated with in-domain and out-of-domain debates.
- We analyse the language of strong and weak arguments according to human scores and (noisy) model predictions. This results in insights into not only which arguments are perceived to be stronger, but also the differences between humans and models. For example, some rhetorical and structural features correlate both with human scores and model predictions, but there are substantial differences in inductive preferences by the model.

2. Related Work

We group the discussion of related work into empirical methods and theoretical models from CAT.

2.1. Empirical Methods for Debate Modeling

While natural language inference and its applications have been extensively studied (Romanov and Shivade, 2018; Kwak et al., 2022; Koreeda and Manning, 2021, inter alia), research that focuses on modeling argumentation graphs is much more limited. Kuhlmann and Thimm (2019) and Craandijk and Bex (2020) used feed forward neural networks and graph neural networks over abstract argument graphs to learn extension-based argumentation semantics. More recently, Al Anaissy et al. (2024) trained various graph neural networks to learn *gradual* semantics, i.e., where argument acceptability degrees take *continuous* values, over argument graphs that contain both attack and support edges. However, all these works rely on graphs containing *abstract* arguments. In contrast, we focus on structured debates where arguments are natural language claims. We show that modeling the text behind these claims is critical for good performance.

Several recent works have explored the interplay between textual and structural cues for stance detection (Pick et al., 2022; Li et al., 2018; Sridhar et al., 2015; Barel et al., 2024). These methods typically address classification tasks, such as predicting pairwise stance relations or assigning discrete polarity labels to arguments, often by leveraging structural embeddings or multimodal representations. Our objective differs both in formulation and in scope: rather than inferring stance, we aim to estimate perceived argumentative relevance by predicting continuous vote distributions at the node level. This entails a shift from polarity detection to credibility modeling, with distinct modeling assumptions and evaluation goals. Crucially, such modeling relies on the availability of argument-level voting signals, which are absent from commonly used stance benchmarks (e.g., 4Forums, ConvinceMe). These corpora, while valuable for stance prediction, do not support the core task investigated here.

Similar to us, Agarwal et al. (2022) model structured debates from Kialo. However, their goal is to infer *pro*/*con* edge labels in debate graphs, a simpler task closer to NLI, rather than predicting the entire distribution of impact votes like we do.

Most recently, Moniri et al. (2024) introduced an automated benchmarking framework based on debates among LLMs to evaluate the LLMs' argumentative reasoning skills. While this differs from our task, it highlights the importance of modeling debates in the LLM landscape.

2.2. Computational Argumentation Theory

The computational argumentation community has adopted a graph-based view of argumentation since the seminal paper by Dung (1995) and has since explored this perspective primarily from a theoretical standpoint. More recently, platforms like Kialo have attracted attention from argumentation researchers (Young et al., 2021; Thieyre et al., 2024). In argumentation theory, debates are typically modeled as graphs, much like in Kialo, where each argument may be assigned an initial weight (e.g., reflecting the level of trust in the source of the argument). The structure of attacks and supports, along with these initial weights, is then used to compute the acceptability degree of each argument (Amgoud et al., 2022).

A compelling question is whether the scores assigned by Kialo users represent initial weights or the final acceptability degrees. We hypothesise that users cast their votes based on their impression of the argument itself, likely before considering the broader structure of the debate. Therefore, these scores are better interpreted as initial weights rather than acceptability degrees. This insight opens the door to significant applications within computational argumentation research. Notably, initial weights are often generated randomly when creating datasets for testing scientific hypotheses or other purposes, as pointed out by Al Anaissy et al. (2024). Our work offers a more realistic and empirically grounded method for estimating such weights, thereby im-

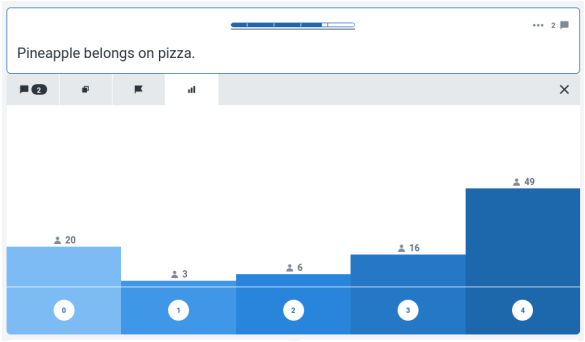

Figure 1: Example of a claim and its corresponding vote distribution (0–4 impact scale: minimal to maximal).

proving the quality of these experimental setups.

3. Dataset

3.1. Source Platform: Kialo

We collect our data (Brouat et al., 2026) from Kialo, a collaborative platform designed for structured debates. Each discussion is organised as an argumentation graph, where nodes represent claims, and directed edges encode argumentative relations, either support or attack (corresponding to the platform's *pro* and *con* stances, respectively). Debates are initiated by a main claim, and users contribute supporting or attacking statements in a tree-like structure, where each node is linked to a unique parent. Each claim consists of a short natural language sentence and is associated with user votes on a 0–4 scale, reflecting its perceived relevance (or impact). This distribution ranges from 0 = *Not at all impactful* to 4 = *Very impactful*. It provides a weak supervision signal for measuring argumentative impact (Figures 1, 2).[2]

Our dataset comprises 181 debate graphs, totalling approximately 83,000 claims, averaging 26 tokens each. The support and attack relations are nearly balanced, with 49.2% support and 50.8% attack edges. Debate trees reach a mean depth of 5.5 ($\pm$1.9), with the deepest spanning 19 levels. The majority of nodes are leaves (60%); non-leaf nodes have on average 2.5 children, with a maximum branching factor of 30.

In compliance with the Kialo license, we do not redistribute debate texts. Instead, we provide the URLs of all debates used, together with scripts to download and preprocess them into graph format. To ensure reliability in the supervision signal, we retained only claims with at least five user votes as training targets, while preserving the entire graph

[2]Kialo's voting system measures the *impact* of a claim, defined as a combination of its veracity and relevance with respect to its parent claim.

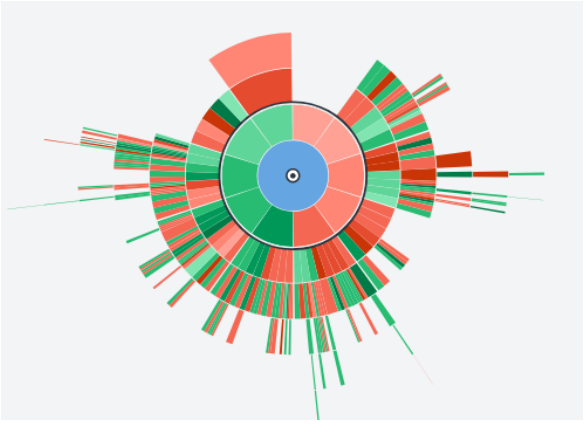

Figure 2: Argumentation tree from the "pineapple on pizza" debate, displayed in a radial layout. The root claim occupies the centre, with each concentric ring corresponding to a deeper level of the debate. Claims in outer rings respond to their parent in the adjacent inner ring, either by supporting it (green edges, *pro*) or attacking it (red edges, *con*).

structure so that all nodes, including unvoted ones, participate in message passing.

3.2. Thematic Grouping

To evaluate the model's capacity to generalise across topical domains, we group a subset of debate graphs into thematic clusters (e.g., ethics, politics, religion). Each debate is represented by the SBERT embedding of its root claim, which captures the framing of the debate in a compact form. We apply k-means clustering over these embeddings, subject to a balancing constraint ensuring that each cluster contains a comparable number of claims with at least five user votes. Not all 181 debates are assigned to a theme; graphs that do not fall cleanly into one of the identified clusters are excluded from this evaluation. Clusters are manually reviewed to ensure thematic coherence, which is feasible given the small number of groups. The resulting grouping enables controlled evaluations of in-domain and out-of-domain transfer. Each theme contains approximately 243 claims with at least five user votes, and on average 8–9 distinct argumentation graphs.

4. Approach

4.1. Task Overview

We address the task of predicting the impact distribution of an argumentative claim from both its textual content and its local context within a debate graph. Formally, let $G = (V, E)$ be a directed graph where each node $v_i \in V$ is associated with a textual sequence x_i and a probability distribution $y_i \in \Delta^4$ over a five-point impact scale. Given a subset of

observed pairs (x_i, y_i), the objective is to learn a function $f_\theta : (x_i, G) \mapsto \hat{y}_i$ that predicts the impact distribution for each node. This constitutes a node-level regression task over text-attributed graphs with sequential input, where the target distribution reflects user-assigned impact scores ranging from 0 (not impactful) to 4 (very impactful).

4.2. Textual Encoding with SBERT

Each claim text x_i is encoded using a transformer-based encoder (SBERT), which maps the input into a dense embedding:

$$s_i = \text{SBERT}_\phi(x_i) \in \mathbb{R}^d,$$

where s_i denotes the embedding of claim x_i and ϕ are the pre-trained (and optionally fine-tuned) parameters. We initialise SBERT with the `paraphrase-MiniLM-L6-v2` model, selected for its lightweight architecture (6 transformer layers), its paraphrase-oriented training objective well-suited to the concise nature of Kialo claims, and its strong sentence-level performance despite its compact size. In most experiments, the encoder is fine-tuned jointly with the downstream components; results with frozen weights are also reported.

To adapt the SBERT embeddings to the input dimensionality of the graph encoder, we apply a feedforward projection:

$$z_i = \text{Proj}(s_i) \in \mathbb{R}^{d'},$$

where Proj is a two-layer FF with ReLU activation. Two layers provide sufficient capacity to re-align the SBERT embedding space with the GAT input space, and were selected based on empirical validation against alternative configurations.

4.3. Graph Modeling with GAT

Given the projected embeddings z_i, we apply a two-layer graph attention network (GAT) to compute context-aware node representations by aggregating information from argumentative neighbors. We use two layers based on theoretical and practical considerations. From a theoretical perspective, impact scores in Kialo are defined relative to a claim's direct parent, suggesting that local context is the primary signal. Two layers allow the model to capture not only direct parent–child relations but also second-order interactions (e.g., whether a claim's parent is itself a support or an attack, and whether a claim's children are themselves supported or attacked), while avoiding over-smoothing and the propagation of potentially irrelevant long-range signals. Exploring deeper architectures would be a natural extension, though this was precluded by the computational demands of the full SBERT×FF×GAT pipeline. The input features are initialised as $\mathbf{h}_i^{(0)} = z_i$, and updated layer-wise according to the following recurrence:

$$\mathbf{h}_i^{(l+1)} = \sigma\left(\sum_{j \in \mathcal{N}(i)} \alpha_{ij}^{(l)} \mathbf{W}^{(l)} \mathbf{h}_j^{(l)}\right),$$

where σ is a non-linear activation function, $\mathbf{W}^{(l)}$ a learnable weight matrix, and $\alpha_{ij}^{(l)}$ an attention coefficient.

Attention weights are computed by:

$$\alpha_{ij}^{(l)} = \text{softmax}_{j \in \mathcal{N}(i)}\left(\mathbf{a}^\top \left[\mathbf{W}^{(l)} \mathbf{h}_i^{(l)} \,\|\, \mathbf{W}^{(l)} \mathbf{h}_j^{(l)}\right]\right),$$

where $\|$ denotes concatenation and $\mathbf{a}$ a learnable feedforward layer with LeakyReLU activation.

The final representation $\mathbf{h}_i^{(2)}$ is passed through a linear layer and softmax classifier to predict the impact distribution:

$$\hat{y}_i = \text{softmax}(W \mathbf{h}_i^{(2)} + b).$$

This architecture enables the model to integrate both semantic content and argumentative structure, capturing how supporting and attacking relations influence the perceived strength of claims.

4.4. Training Details

Our task involves predicting full probability distributions over the 5-point impact scale. While prior work has suggested that simple regression losses such as mean squared error (MSE) or mean absolute error (MAE) can outperform Kullback–Leibler divergence (KLDiv) in tasks with soft labels (Müller et al., 2019; Reimers and Gurevych, 2019), we conducted a controlled comparison between these objectives. In addition to standard MSE, we evaluated restricted variants that consider only the top-2 or top-3 classes of the target distribution. Empirically, KLDiv consistently yielded the best performance.

This outcome is consistent with the nature of our data. Vote distributions in Kialo debates are often polarised or sharply peaked, reflecting disagreements among users. In such cases, MSE encourages predictions close to the mean, which can obscure underlying divergences. KLDiv, by contrast, rewards preservation of the full distributional shape, making it more appropriate for modelling argumentative impact. As an illustration, Figure 1 shows a bimodal distribution of votes for a claim about pineapple on pizza, a canonical example of a polarising topic.

5. Experiments

Models are trained to minimise the KLDiv loss between predicted and gold vote distributions, as motivated in the previous section. For evaluation, however, we report the MAE on the mean of these distributions. The mean score is a single scalar on the

Model	MAE (all claims)	MAE (top claim)
Baselines		
SBERT×MLP	0.105 ± 0.005	0.096 ± 0.001
LLM (restricted graph)	–	0.2420
LLM (full prompt)	–	0.2609
Our method		
SBERT×FF×GAT	0.068 ± 0.001	0.061 ± 0.001
Ablations		
SBERT×GAT	0.083 ± 0.002	–
SBERT×FF×GCN	0.085 ± 0.001	–
FF×GAT	0.110 ± 0.005	–

Table 1: Mean absolute error (MAE) obtained by different models on the entire test partition. Standard deviation reported over five runs with different random seeds when available. We report MAE values for all claims in a debate as well as for the top claim only.

0–4 scale and thus has the same dimensionality as CAT's initial weights and (scalar) acceptability degrees, which makes the error directly interpretable within argumentation theory. Each debate graph is assigned in its entirety to either training or test, preventing structural leakage. Only claims with at least five votes are used as supervision targets, but all nodes participate in message passing and gradient updates.

5.1. What Does the Model Learn?

Debate Structure Matters

We compare our full pipeline (SBERT×FF×GAT) with a textual baseline (SBERT×MLP). The former integrates SBERT embeddings, a feedforward (FF) compression layer, and a GAT to model attack and support relations. The latter directly maps SBERT embeddings to vote distributions via a multi-layer perceptron (MLP). Adding an intermediate FF layer to the MLP baseline did not improve performance and was therefore omitted. This experiment highlights that integrating argumentative structure leads to a marked performance gain: the MAE drops from 0.105 to 0.068 from the SBERT×MLP to our method (Table 1). This experiment demonstrates that the graph encodes information that cannot be recovered from the text alone.

Also, unlike the text-only model, our GAT-based architecture benefits from an intermediate dimensionality reduction via a feed-forward layer (FF in Table 1). The slight compression of textual information is offset by the graph attention network's ability to propagate and enrich representations through structured neighbourhoods. Indeed, aggregating over pre-synthesised representations proves more effective than operating directly in the high-dimensional SBERT space.

Claim Text also Matters

While the previous paragraph highlights the contribution of structure, we must ensure it is not the sole source of predictive signal. In our complete model (SBERT×FF×GAT), SBERT encodes each claim into a dense vector, which serves as input to the FF, GAT and prediction layers. To isolate the role of textual content, we remove SBERT and inject fixed random vectors of the same size. These vectors are sampled once, held constant throughout training, and contain no linguistic information. The rest of the architecture, including the FF, GAT and prediction layers, remains unchanged. Despite this, performance drops by approximately 0.04 MAE (61.78%), suggesting that the semantic representations produced by SBERT play a substantial role in prediction. We further analyse the linguistic signals underlying this contribution in Section 6.

Textual Representations Adapt to Argumentative Structure

To explore whether the model internalises a form of argumentative language, we compare two variants: one in which SBERT is frozen after fine-tuning on our corpus, and another in which SBERT is jointly trained end-to-end with the rest of the pipeline. Interestingly, the frozen model, which starts from a better-informed representation, reaches a MAE of 0.080, while the trainable variant achieves 0.068. This gap suggests that the model benefits from adapting its textual encoding to the graph-based reasoning task, learning representations that go beyond static semantics and align with the structure of argumentative discourse; in other words, it is learning a structured language of argumentation grounded in graph topology.

Train Theme	Test Theme	SBERT×FF×GAT (ours)	SBERT×MLP (baseline)
Discrimination	Discrimination (ID)	**0.1154**	0.1244
	Religion	0.1930	**0.1888**
	Animal ethics	**0.1202**	0.1298
	Politics	**0.0903**	0.1208
	Governance	**0.1335**	0.1623
Religion	Discrimination	**0.1060**	0.1332
	Religion (ID)	0.1812	**0.1303**
	Animal ethics	**0.1108**	0.1170
	Politics	**0.0923**	0.1065
	Governance	**0.1402**	0.1487
Animal ethics	Discrimination	**0.1118**	0.1460
	Religion	**0.1877**	0.1894
	Animal ethics (ID)	0.1206	**0.0730**
	Politics	**0.0882**	0.1147
	Governance	**0.1473**	0.1614
Politics	Discrimination	**0.1083**	0.1475
	Religion	0.1906	**0.1876**
	Animal ethics	**0.1122**	0.1211
	Politics (ID)	**0.0858**	0.1065
	Governance	**0.1399**	0.1572
Governance	Discrimination	**0.1102**	0.1463
	Religion	0.1942	**0.1867**
	Animal ethics	**0.1139**	0.1237
	Politics	**0.0846**	0.1078
	Governance (ID)	0.1434	**0.1084**
Macro avg (ID)		0.1293	**0.1085**
Macro avg (OOD)		**0.1288**	0.1448

Table 2: Cross-domain evaluation: MAE for our method (SBERT×FF×GAT) and baseline (SBERT×MLP) for all theme combinations in Kialo.

5.2. Debate Structure Increases Generalisation Capabilities

Debates naturally span a diverse range of topics, yet argumentation theory suggests that the art of debate relies on structural and rhetorical patterns that should transcend thematic boundaries (Dung, 1995). If our model truly captures these general principles, as suggested by its reliance on discourse structure (Section 5.1) and its ability to adapt semantic encodings (Section 5.1), then it should be able to transfer across topics. To test this hypothesis, we carry out cross-domain evaluations, training the model on one thematic cluster and assessing its performance on the others, following the grouping introduced in Section 3.2. This setup allows us to test how well a model trained in one topical domain transfers to another.

Our results (Table 2) show that while the purely textual baseline (SBERT×MLP) often performs well in-domain, it degrades significantly when transferred to new themes, which indicates overfitting. In contrast, our full model (SBERT×FF×GAT) displays more stable behaviour and consistently lower MAE in out-of-domain settings. The average out-of-domain MAE drops from 0.1448 with the baseline to 0.1288 with our model. This suggests that the structural features of argument graphs encode domain-agnostic regularities in argumentative discourse, enabling the model to generalise more effectively across thematic boundaries. These findings support our hypothesis that the inclusion of graph structure improves cross-domain robustness.

A note on structural imbalance. An interesting exception arises in the *Religion* theme, observed in the intra-theme experiment where training is restricted to a single domain. Here, our model underperforms the text-only baseline in cross-domain transfer. This effect is not a general weakness of structure-aware models but rather a consequence of the particular composition of this subset: it consists almost entirely of a single large debate tree, included so that the theme would contain a sufficient number of high-quality claims (i.e., those with 5 votes or more) to meet the standards applied

to other themes, which typically comprise multiple smaller debate graphs.

While GAT are, in principle, robust to graph size due to local message passing, the topology of the training data plays a critical role in cross-domain transfer. When trained on a single large debate tree (as in the Religion theme), high-quality signals are concentrated near the root, which encourages the model to internalise debate-specific patterns. Applied to themes composed of multiple smaller debates, this leads to poor transfer: training on a single large tree biases the model towards propagating information along highly centralised branches, a structural pattern absent in forests. In contrast, training on a forest distributes informative signals across many subgraphs, exposing the model to more varied propagation dynamics and improving transfer robustness. This observation highlights how experimental restrictions to structurally imbalanced domains can reveal important sensitivities of structure-aware models to graph topology.

5.3. Evaluating LLMs via In-Context Learning on Debate Roots

When it comes to processing textual information, we were naturally curious about how a large language model like *LLaMA-4-Maverick* (2025) would perform on our task. Predicting the impact of all claims across full debate graphs quickly proved infeasible due to prompt size and token limits. However, focusing on a single prediction, namely, that of the main claim, offers a practical and meaningful alternative. As the central node, the main claim frames the discussion and typically receives the most attention. Its predicted score thus provides a synthetic signal of the overall argument perception. Since only one output is required, the full textual and structural context can be included without exceeding input constraints. The main claim's vote distribution is masked, letting the model reason over the graph before outputting its judgment.

We prompt LLaMA-4-Maverick to return a probability distribution over Kialo's five-level impact scale. This model was chosen for its general performance and ability to process long sequences, essential for encoding debates in textual form. The prompt preserves support and attack relations, with the main claim marked explicitly. We refer the reader to Appendix A for the exact prompt template and decoding parameters.

Despite these favourable conditions, performance remains poor: the MAE reaches 0.2609, barely above the 0.2530 obtained by always predicting a uniform distribution. A reduced graph version, limited to the first three dialogue layers, does slightly better (0.2420), suggesting that immediate context matters more than structural depth, or that the LLM simply struggles to process large graph structures. In contrast, our supervised model (SBERT$\times$FF$\times$GAT) achieves an MAE of 0.061 on the same task, underscoring its superior ability to integrate semantics and structure, and illustrating the limitations of current LLMs for fine-grained argumentative reasoning. We regard this experiment as a preliminary probe rather than a strong baseline, designed to illustrate the current limitations of off-the-shelf LLMs when applied to argumentative impact prediction. It also motivates our ongoing work on the integration of argumentative structure into LLM-based approaches.

6. Discussion and Further Analyses

We examine linguistic and structural features associated with impact scores, for both human judgments and model predictions, highlighting where they converge and where they diverge.

6.1. Overview of Linguistic and Structural Feature Effects

Beyond graph-based reasoning, we conduct an exploratory analysis of several linguistic and structural features in order to better understand the mechanisms underlying both human credibility judgments and model predictions. Our goal is not to build a predictive model from these features, but rather to assess whether certain surface properties correlate with how arguments are evaluated.

Our selection of linguistic features follows prior work in Argument Quality Assessment (AQA) and persuasion studies, which have discussed rhetorical and stylistic markers as potential indicators of argument quality or persuasiveness. In particular, Wachsmuth et al. (2017) survey theoretical and practical perspectives on argument quality, while Wachsmuth et al. (2024) revisit the notion of quality in the context of recent large language models. Complementarily, Tan et al. (2016) provide empirical evidence that markers such as hedging can contribute to persuasion in online discussions. Motivated by these insights, we focus on categories such as hedging, emotional language, pronouns, and negation, operationalised through established lexicons adapted from prior work (Johns, 2001; Farkas et al., 2010; Morante and Blanco, 2012; Mohammad and Turney, 2013; Davidson et al., 2017). We also include simple textual measures (number of characters and tokens), social markers (first- and second-person pronouns), and graph-related properties (number of supporting or attacking children, attention from the parent node).

Each feature is analysed using a univariate analysis of variance (ANOVA, Fisher, 1925), with both the human score and the model score as dependent

	Human (gold) score		Model (predicted) score	
Feature	η^2	p-value	η^2	p-value
Num_Chars	0.0001	5.8×10^{-2}	0.0013	$\mathbf{9.5 \times 10^{-14}}$
Num_Tokens	0.0001	$\mathbf{3.8 \times 10^{-2}}$	0.0015	$\mathbf{1.1 \times 10^{-15}}$
Negation	0.0000	4.9×10^{-1}	0.0008	$\mathbf{2.7 \times 10^{-9}}$
Hedging	0.0002	$\mathbf{5.0 \times 10^{-3}}$	0.0002	$\mathbf{3.0 \times 10^{-3}}$
Offensive	0.0000	4.7×10^{-1}	0.0001	$\mathbf{3.2 \times 10^{-2}}$
Emotion	0.0001	$\mathbf{1.5 \times 10^{-2}}$	0.0001	1.1×10^{-1}
Pronoun_I	0.0001	7.1×10^{-2}	0.0000	4.0×10^{-1}
Pronoun_You	0.0003	$\mathbf{9.0 \times 10^{-4}}$	0.0000	4.8×10^{-1}
Pronoun_We	0.0003	$\mathbf{2.0 \times 10^{-4}}$	0.0000	3.9×10^{-1}
Pronoun_They	0.0000	8.4×10^{-1}	0.0001	1.1×10^{-1}
Num_Support	0.0001	$\mathbf{1.7 \times 10^{-2}}$	0.0024	$\mathbf{1.7 \times 10^{-24}}$
Num_Attack	0.0004	$\mathbf{8.1 \times 10^{-5}}$	0.0015	$\mathbf{7.0 \times 10^{-16}}$
Parent_Attention	0.0014	$\mathbf{4.7 \times 10^{-15}}$	0.0008	$\mathbf{1.7 \times 10^{-9}}$

Table 3: ANOVA analysis of linguistic and structural features. For each feature, we report η^2 (effect size) and p-value for both human and model score. Bold values indicate statistical significance ($p < 0.05$).

variables. Intuitively, the former analysis measures how much the feature influences human credibility judgments, while the latter indicates what the model learns. Alongside the p-value, we report the effect size η^2, which estimates the proportion of variance in the target variable explained by the feature. This analysis provides an initial lens into factors that may influence perceived credibility and highlights the potential biases or inductive signals exploited by the model.

6.2. Features Influencing Human Judgments

Indeed, as shown in Table 3, several features exhibit statistically significant effects on the *Human score*, as derived from human votes. Among them are the use of second-person `you` and first-person plural `we` pronouns, as well as the presence of hedging and emotional language. Structural attention from the parent node also emerges as a relevant factor. While these effects are statistically significant ($p < 0.05$), the corresponding η^2 values remain small, typically below 0.001, indicating that each feature accounts for only a marginal proportion of the variance. Nonetheless, their significance suggests that human judgments are not indifferent to certain rhetorical markers or to the local structure of the debate graph.

6.3. Features Influencing Model Predictions

Several features exhibit statistically significant effects on the *Model score*, as output by our model. These include basic textual metrics such as the number of tokens and characters, the presence of hedging and negation, as well as structural indicators like the number of supporting children and the attention weight from the parent node.

As with human judgments, these effects are statistically significant ($p < 0.05$) but small in magnitude: most η^2 values remain under 0.002, indicating limited explanatory power. Still, the model's sensitivity to these surface-level and local structural cues suggests that it internalises shallow yet consistent patterns from the training data, aligning its predictions with frequently occurring syntactic forms and graph configurations.

6.4. Weak Linguistic Correlations and Model–Human Divergences

Among the analysed features, *parental attention* stands out as a common driver for both humans and the model, reflecting the importance of local argumentative structure. Beyond this shared reliance, we observe divergences. Features such as *second-person pronouns* influence human judgments but not the model, whereas *negation* is exploited by the model despite showing no clear effect on human votes. These discrepancies are not straightforward to interpret, yet a consistent trend emerges: surface linguistic cues highlighted in prior AQA work emerge as statistically significant for the model, while their influence remains comparatively weaker for humans. In both cases, structural features exert a stronger and more stable effect, consistent with the way argumentative relations are organised in multi-branch debate graphs.

While the observed effects of individual linguistic markers are weak in absolute terms, they are statistically significant. This is precisely the purpose of the ANOVA analysis: to identify subtle yet consistent signals, rather than strong predictors in isolation. The relatively small effect sizes therefore

do not undermine the analysis; rather, they highlight a key contrast with prior Argument Quality Assessment or persuasion studies, where such features (e.g., hedging, emotionality) were found to play a stronger role. Our discussion below explains why these effects appear attenuated in our setting.

A likely contributor to the weak overall correlations is the nature of the Kialo platform, where moderation encourages polite, rational discourse. This yields stylistically uniform texts, making it harder to isolate linguistic cues that separate strong from weak arguments. At the same time, Kialo provides a uniquely rich testbed: it is, to our knowledge, the only large-scale platform combining explicit support and attack relations, claim-level impact annotations, and debates spanning diverse and often polarising topics across many domains. This combination ensures broad lexical coverage while maintaining stylistic neutrality, which attenuates the predictive power of isolated surface features such as hedging or emotionality. Moreover, as shown in Section 5.1, our model adapts its textual representations to the graph-based task, learning embeddings shaped by argumentative role and position. In such a context, language becomes structurally mediated, and correlations based on independent linguistic markers are expected to be attenuated.

7. Conclusion

This work explored the capacity of neural models to capture the structure and content of argumentative discourse. We introduced a graph-based architecture that jointly models textual semantics and relational structure, integrating contextualised claim embeddings with graph attention mechanisms. Evaluated on a large set of structured debates, the model achieved strong predictive performance and demonstrated robust out-of-domain generalisation. An analysis of its learned representations further revealed a sensitivity to rhetorical and structural signals, some of which mirror human judgments, while others reflect distinct inductive biases of the model.

8. Ethical considerations and limitations

Our training data originates exclusively from the Kialo platform, which remains, to our knowledge, the only large-scale source providing explicit support and attack relations together with claim-level impact annotations. While this makes Kialo a uniquely valuable testbed for modelling structured argumentation, its moderated and self-selected community may not reflect the full diversity of argumentative styles found in other domains.

The model learns to approximate collective human judgments of impact, which are inherently social and context-dependent. Predictions should therefore be interpreted as estimates of perceived argumentative relevance rather than objective measures of quality or truth, particularly in evaluative or decision-making contexts.

Acknowledgements

The authors benefited from the support of the joint CNRS – University of Arizona PhD programme, project SURFING. Maxime Brouat and Srdjan Vesic also benefited from the support of the project AGGREEY ANR-22-CE23-0005 of the French National Research Agency (ANR).

Bibliographical References

Vibhor Agarwal, Sagar Joglekar, Anthony P Young, and Nishanth Sastry. 2022. Graphnli: A graph-based natural language inference model for polarity prediction in online debates. In *Proceedings of the ACM Web Conference 2022*, pages 2729–2737.

Caren Al Anaissy, Sandeep Suntwal, Mihai Surdeanu, and Srdjan Vesic. 2024. On learning bipolar gradual argumentation semantics with neural networks. In *Proceedings of the 16th International Conference on Agents and Artificial Intelligence, ICAART 2024, Volume 2, Rome, Italy, February 24-26, 2024*, pages 493–499. SCITEPRESS.

Leila Amgoud, Dragan Doder, and Srdjan Vesic. 2022. Evaluation of argument strength in attack graphs: Foundations and semantics. *Artif. Intell.*, 302:103607.

Guy Barel, Oren Tsur, and Dan Vilenchik. 2024. Acquired taste: Multimodal stance detection with textual and structural embeddings.

Pietro Baroni, Dov Gabbay, Massimilino Giacomin, and Leendert van der Torre, editors. 2018. *Handbook of Formal Argumentation*. College Publications.

Dennis Craandijk and Floris Bex. 2020. Deep learning for abstract argumentation semantics. *arXiv preprint arXiv:2007.07629*.

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *CoRR*, abs/1703.04009.

Phan Minh Dung. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artif. Intell.*, 77(2):321–358.

Richárd Farkas, Veronika Vincze, György Móra, Janos Csirik, and György Szarvas. 2010. The conll-2010 shared task: Learning to detect hedges and their scope in natural language text. *Proceedings of the Fourteenth Conference On Computational Natural Language Learning (CoNLL-2010): Shared Task*, pages 1–12.

Ronald A. Fisher. 1925. *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh. First edition; later expanded in multiple reprints.

Ann Johns. 2001. Hedging in scientific research articles: Ken hyland. amsterdam/philadelphia: John benjamins publishing co., box 75577, 1070 an amsterdam, the netherlands, 1998, 309 pp. *English for Specific Purposes*, 20:200–203.

Yuta Koreeda and Christopher Manning. 2021. ContractNLI: A dataset for document-level natural language inference for contracts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1907–1919, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Isabelle Kuhlmann and Matthias Thimm. 2019. Using graph convolutional networks for approximate reasoning with abstract argumentation frameworks: A feasibility study. In *International Conference on Scalable Uncertainty Management*, pages 24–37. Springer.

Alice S. Kwak, Jacob O. Israelsen, Clayton T. Morrison, Derek E. Bambauer, and Mihai Surdeanu. 2022. Validity assessment of legal will statements as natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP 2022*.

Chang Li, Aldo Porco, and Dan Goldwasser. 2018. Structured representation learning for online debate stance prediction. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3728–3739, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *CoRR*, abs/1308.6297.

Behrad Moniri, Hamed Hassani, and Edgar Dobriban. 2024. Evaluating the performance of large language models via debates. *arXiv preprint arXiv:2406.11044*.

Roser Morante and Eduardo Blanco. 2012. Semantic scope of negation: Review and proposal for an annotation scheme. *Computational Linguistics*, 38(3):281–321.

Rafael Müller, Simon Kornblith, and Geoffrey E. Hinton. 2019. When does label smoothing help? *CoRR*, abs/1906.02629.

Ron Pick, Vladyslav Kozhukhov, Dan Vilenchik, and Oren Tsur. 2022. Stem: Unsupervised structural embedding for stance detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36:11174–11182.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, abs/1908.10084.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from natural language inference in the clinical domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1586–1596, Brussels, Belgium. Association for Computational Linguistics.

Mobashir Sadat and Cornelia Caragea. 2022. SciNLI: A corpus for natural language inference on scientific text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7399–7409, Dublin, Ireland. Association for Computational Linguistics.

Henri Savigny and Bruno Yun. 2025. Amelia: A family of multi-task end-to-end language models for argumentation. In *Proceedings of the 2025 Annual Meeting of the Association for Computational Linguistics (ACL)*, Vienna, Austria. Association for Computational Linguistics. ArXiv preprint arXiv:2508.17926.

Dhanya Sridhar, James Foulds, Bert Huang, Lise Getoor, and Marilyn Walker. 2015. Joint models of disagreement and stance in online debate. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 116–125, Beijing, China. Association for Computational Linguistics.

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)*, pages 613–624. International World Wide Web Conferences Steering Committee.

Jordan Thieyre, Aurélie Beynier, Nicolas Maudet, and Srdjan Vesic. 2024. Reassessing the impact of reading behaviour in online debates under the lens of gradual semantics. In *Proceedings of the Fifth International Workshop on Systems and Algorithms for Formal Argumentation co-located with 10th International Conference on Computational Models of Argument (COMMA 2024), Hagen, Germany, September 17th, 2024*, volume 3757 of *CEUR Workshop Proceedings*, pages 119–133. CEUR-WS.org.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks.

Henning Wachsmuth, Khalid Al Khatib, Yamen Ajjour, and Johannes Kiesel. 2024. Argument quality assessment in the age of instruction-following large language models. In *Proceedings of the 2024 International Conference on Language Resources and Evaluation (LREC)*, pages 1254–1264, Torino, Italy. European Language Resources Association (ELRA).

Henning Wachsmuth, Nona Naderi, Ivan Habernal, Yufang Hou, Vinodkumar Prabhakaran, Graeme Thijm, Graeme Hirst, and Benno Stein. 2017. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 176–187, Valencia, Spain. Association for Computational Linguistics.

Anthony P. Young, Sagar Joglekar, Gioia Boschi, and Nishanth Sastry. 2021. Ranking comment sorting policies in online debates. *Argument Comput.*, 12(2):265–285.

Language Resource References

Maxime Brouat, Mihai Surdeanu, Srdjan Vesic, and Eduardo Blanco. 2026. Kialo debate links dataset. `https://github.com/arg-ml/debate-impact-modeling/blob/main/kialo_links.txt`. Text file listing the URLs of all debates used in this study. Provided to comply with Kialo's license, which permits linking but not redistribution of debate content.

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *CoRR*, abs/1703.04009.

Richárd Farkas, Veronika Vincze, György Móra, Janos Csirik, and György Szarvas. 2010. The conll-2010 shared task: Learning to detect hedges and their scope in natural language text. *Proceedings of the Fourteenth Conference On Computational Natural Language Learning (CoNLL-2010): Shared Task*, pages 1–12.

Ann Johns. 2001. Hedging in scientific research articles: Ken hyland. amsterdam/philadelphia: John benjamins publishing co., box 75577, 1070 an amsterdam, the netherlands, 1998, 309 pp. *English for Specific Purposes*, 20:200–203.

Meta AI / Meta Llama team. 2025. *LLaMA-4 Maverick (17B, 128E) Instruct / Multimodal model*. PID https://huggingface.co/meta-llama/Llama-4-Maverick-17B-128E-Instruct. Mixture-of-Experts model with 128 experts, supports up to 1 M token context window.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *CoRR*, abs/1308.6297.

Roser Morante and Eduardo Blanco. 2012. Semantic scope of negation: Review and proposal for an annotation scheme. *Computational Linguistics*, 38(3):281–321.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *CoRR*, abs/1908.10084.

A. Data, Lexicons, and Prompt Template

A.1. Debate Links and Data Access

In accordance with the Kialo license, we do not redistribute the debate content directly. Instead, we provide a file `kialo_links.txt` in the repository, which contains the URLs of all Kialo debates used in our experiments. Each link points to a public debate that served as source data for our work.

A.2. Lexicons

We provide the lexicons used to compute hedging, negation, offensive language, and emotional content features. The file `lexicon.txt`, included in the repository, contains one entry per line, grouped by category. These lists are adapted from existing resources: hedging from Johns (2001); Farkas et al. (2010), offensive terms from Davidson et al. (2017), emotions from Mohammad and Turney (2013), and negation from Morante and Blanco (2012).

A.3. Prompt Template for LLM Evaluation

To evaluate the ability of a large language model to predict the credibility distribution of a main claim,

we designed the following prompt. It introduces the Kialo debate structure, defines the voting scale, and requests a numerical prediction without explanation. The model used in this evaluation was `llama-4-maverick`, a variant of LLaMA-4 accessed via the OpenRouter API. The prompt was provided alongside a debate graph formatted as a sequence of claims and argumentative relations.

> You are analyzing argumentation graphs from Kialo, a debate platform where users vote on the impact of claims, defined as a combination of their veracity and relevance with respect to their parent claim.
> - Each claim is connected by supports (reinforcing the parent claim) or attacks (challenging the parent claim).
> - Votes are given as a distribution over five categories: [not impactful, slightly impactful, moderately impactful, very impactful, extremely impactful].
> - Some claims have votes [0,0,0,0,0], which means no votes have been cast on them yet. This does not mean the arguments are invalid, just that they lack voting data.
>
> Main Objective: Predict the vote distribution for the main claim (shown as [MASKED]).
>
> Output Format: Provide 5 numbers summing to 1, corresponding to [not impactful, slightly impactful, moderately impactful, very impactful, extremely impactful]. Do not explain your reasoning, only output the 5 numbers in a list.