Argument Move Classification

Lachlan Wallbridge	Dhruv Joshi	Arkar Myo	Myles Benyon	Farhan Raonak
z5359327	z5448031	z5459275	z5363307	z5481648

ABSTRACT

Online debate analysis has become increasingly critical for understanding public discourse and preventing spreading misinformation. This paper is a summary of the improved version of the model called GraphNLI which is a novel approach to argument move classification that leverages graph structure through strategic walk sequences and optimized sentence transformers. The dataset used to train these models are 1560 threads from Kialo debate platform.

We introduced weighted root-seeking walks with exponential decay (0.75 factor) to capture the broader argument contexts demonstrating relevant information from distant nodes. In this approach, different kinds of sentence transformer models are being used to train and the DeBERTa-v3-xsmall (22M parameters) achieves optimal performance-efficiency balance, outperforming larger base models including the GraphNLI baseline (82.87% accuracy) while maintaining computational feasibility. The final accuracy reached 83.14% with 0.9020 ROC AUC establishing a new benchmark for parameter-efficient argument classification.

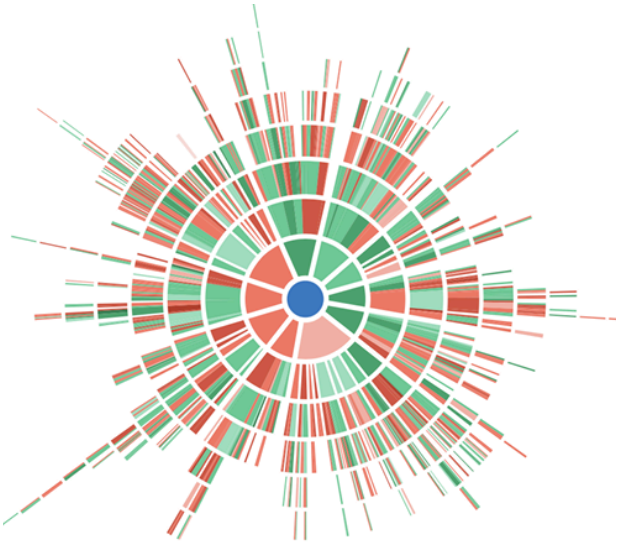
KEYWORDS

Online Debates, Argument Mining, Polarity Prediction, Kialo, GraphNLI, Sentence Transformers, DeBERTa,

I. INTRODUCTION

Debating is a precious resource that allows us as a society to be able to function more efficiently. By sharing ideas, understanding viewpoints and learning from others, we can work together to achieve tasks that are larger in scope than what we can handle alone. Thus by partaking in debating, we better understand everyone and form well thought out opinions and ideals that can be thoroughly backed up with evidence and proof.

The internet is a prime catalyst to allow debates from all over the world to occur and thus users with differing backgrounds can communicate with others, leading to many ideas of varying importance to be discussed on a global level. By having debates at this level, common ground and understanding can be achieved with everyone globally which leads to a more unified society that can achieve greater feats through this understanding brought about through debating. However, debates can just as easily devolve into hate fueled speech and misinformation that does not contribute to a healthy discussion. These soured exchanges of ideas do more harm as they can reinforce echo chambers and cause a degenerative change of opinions that other healthy discussions have worked hard to disprove. As these debates must be logged and recorded, it is more important than ever to properly model these debates and classify arguments so that global opinions and stances can be understood. By being able to model debates and discussions, problems like hate speech, false claims and skewed statistics become more apparent and easier to remove from the discussion. Leading to healthier discussions where ideas are properly expressed and understood.



Visualization of the complexity of a debate tree

An important step to reaching this goal is to be able to design a model which can accurately predict the move an argument is making (for or against a given claim). This polarity prediction task would allow us to be able to have live debates where each side's ideas can be clearly understood. Or to generate polls from a topic to show where the public consensus on a topic resides.

We contribute to the field by adapting and extending the GraphNLI framework to create a more efficient and accurate model for polarity prediction in online debates. Our approach combines weighted root-seeking graph walks, an optimised weighted aggregation method, and the DeBERTa-v3-xsmall sentence transformer backbone to capture richer debate context with lower computational cost. It outperforms the original GraphNLI model despite using fewer parameters, showing that architecture and trained vocabulary outweighs raw parameter count in performance impact.

II. RELATED WORK

Graph-NLI

The Graph-NLI model seeks to utilise the tree structure found within the Kialo dataset as a Bipolar Argumentation Framework. This is a framework which sees arguments tied to claims as nodes and edges, similar to the structure found in Kialo debates. Previous works have mainly looked at inputting only the argument and the claim which is enough to classify a basic argument, however the Graph-NLI model seeks to utilise nearby neighbours in the Bipolar Argumentation Framework to achieve further context on the argument. The result of the study was a significantly higher accuracy than previous baselines, which resulted from the additional context gathered by utilising graph walks. It was noted that the context from the parent and grandparent and so on, provided more relevant context than siblings and children. However the importance of the ancestor nodes decreases the further they are from the given argument node. Graph-NLI was trained on a highly moderated platform (Kialo) where the arguments given are clearly defined as for or against. Therefore the validity of this model on more unmoderated platforms like Reddit or Twitter is unknown.

Model	Accuracy (%)
Bag-of-Words + Logistic Regression	67.00
Prompt Embeddings + Logistic Regression	61.20
Sentence-BERT with classifier layer	79.86
BERT Embeddings: Root-seeking Graph Walk + MLP	70.27
GraphNLI: Root-seeking Graph Walk + Sum	80.70
GraphNLI: Root-seeking Graph Walk + Avg.	81.96
GraphNLI: Root-seeking Graph Walk + Weighted Avg.	**82.87**
GraphNLI: Biased Root-seeking Random Walk + Sum	79.95
GraphNLI: Biased Root-seeking Random Walk + Avg.	80.44

Sentence-BERT

Sentence-BERT is a modified version of the BERT network and uses siamese and triplet networks to derive meaning from the argument semantically. In Sentence-BERT the triplet network is used to update the weights to produce semantically meaningful sentence embeddings. This is done by minimising the distance between the anchor and the positive case while maximising the distance between the anchor and the negative case, leading to a clear differentiation between a for or against argument. A pooling operation was also added to the output of the BERT model to produce a fixed sized sentence embedding. The difference between Sentence-BERT and normal BERT was that BERT mapped sentences to a vector space in an inefficient way which led to similarity comparisons to be computationally expensive.

Sentence-BERT, which used siamese and triplet networks to map sentences into a vector space where semantically similar sentences remained close, outperformed BERT by enabling far less computationally expensive similarity comparisons without sacrificing accuracy, which can be accredited to its optimised vector space mapping.

III. METHODS & MODEL

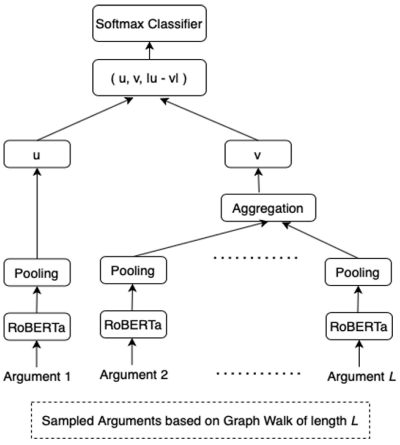
Our model frames polarity classification in argumentative dialogues as a binary edge classification task, using 1560 threads from the Kialo debate trees. Each node represents an argument (including associated metadata such as IDs, timestamps, and text), and each directed edge encodes the stance of the child argument towards its parent: +1 for supporting and -1 for attacking.

To incorporate structural information into a sequence model, debate trees are transformed into short graph walk sequences that preserve contextual relationships. We implement the following two walk strategies:

1. Biased root-seeking random walk (75% probability of moving toward the parent and 25% toward the children).
2. Deterministic weighted root-seeking walk (strictly follows the parent links)

Our approach builds on GraphNLI's "current vs. context" methodology, but modifies both the encoder and aggregation stages. Each sentence is embedded using a Sentence Transformer backbone and applies mean pooling over the token embeddings. The classification head (SoftmaxLoss) takes the following inputs:

- v: which is the vector for the current argument
- u: which is the aggregated vector of the remaining walk positions (ancestors).



We evaluate the following three aggregation strategies:

1. **Mean**: Highlights the average context of the argument chain and features equal weight for all nodes.
2. **Sum**: Aggregates the argument embeddings by accumulating their content.
3. **Weighted**: Implements an exponential decay of importance on the distance for arguments. The rate of decay per node hop is 0.75. This method prioritises closer ancestors while down-weighting distant ones.

Reproducibility is ensured through version-controlled code, fixed seeds, archived logs and exported model artifacts. We focus solely on the Kialo dataset due to its relatively low noise compared to other debate platforms.

IV. EXPERIMENTS

We designed our experiments to isolate the effects of:

1. Amount of graph context fed into the classifier
2. Sentence transformer's ability to convert that context into useful features.

Debate trees from the Kialo dataset are converted into weighted root-seeking walks of length 4 (5 sentences total), with the label taken from the first traversed edge. Texts are tokenised to a maximum of 256 tokens, with shorter walks padded by nulls that are ignored by the aggregation step. This yields 324373 samples, split 80/20 into 259499 training and 64874 development examples, with a mild class imbalance of 57.4% attack against 42.6% support.

To identify the optimal variables so to develop a final model, we first compared differing sentence transformers by fixing mean aggregation and training identical pipelines for the following architectures:

- all-MiniLM-L6-v2 (22.7M PMs, 30.5K Vocab)
- DistilRoBERTa-base (82.8M PMs, 50K Vocab)
- DeBERTa-v3-small (44M PMs, 128K Vocab)
- DeBERTa-v3-xsmall (22M PMs, 128K Vocab)

All pipelines used a batch size of 16, 4 epochs, a learning rate of 5e-5, fp16 AMP, a seed of 42, and pinned dependencies to ensure reproducibility.

We then compared the aggregation methodologies by holding the sentence transformer fixed. The specific aggregation methods compared were mean, sum and weighted, where weighted featured a hop decay factor of 0.75. Ablation and sensitivity tests were implemented for the following research areas to assure model confidence:

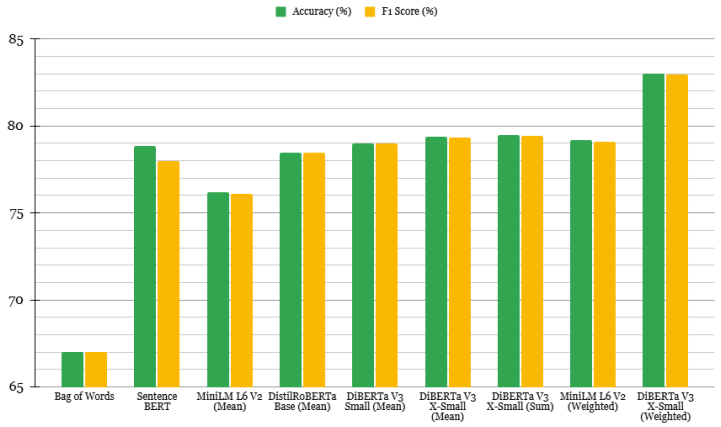
1. Walk length sensitivity (> or < 4-walk)
2. Decay factor sensitivity (> or < 0.75)
3. The removal of |u-v| within the Softmax head
4. The increase of maximum sequence length beyond 256 tokens

During the training and evaluation process, periodic development set checks were run every 21000 steps and once each epoch was completed. Both the sentence transformer backbone and classifier head were saved to ensure consistency between training and evaluation and final evaluations used the following metrics:

- **Accuracy:** Percentage of correct predictions
- **F1-Score:** Harmonic mean of Precision & Recall
- **Precision:** Percentage of true positives against predicted positives
- **Recall:** Percentage of true positives against actual positives
- **ROC AUC:** Models ability to distinguish between classes (score 1: perfect classifier)
- **Confusion Matrix:** Displayed prediction results

V. RESULTS

The following bar chart displays the results of all developed models:



To establish baselines, we first replicated the baseline models presented in the GraphNLI literature:

1. **Bag of Words**: A fundamental model that disregards grammar, word order and context.
2. **Sentence-BERT**: A more sophisticated baseline that generates semantic sentence embeddings using pooling to produce fixed-size vectors.

These baselines performed as expected, providing reference points for evaluating our own architectures.

Analysing the four sentence transformer results that were developed with mean aggregation, highlights that DeBERTa-v3-xsmall marginally outperformed the others, achieving 0.35% higher accuracy and 0.34% higher F1 score than the next-best model.

Despite having the fewest parameters among the tested BERT variants, DeBERTa-v3-xsmall outperformed the larger transformers; DeBERTa-v3-small and DistilRoBERTa-base (the GraphNLI backbone). This demonstrates that architecture and vocabulary coverage can outweigh raw parameter count in performance impact. DeBERTa-v3-xsmall utilises an architecture of 12 layers with size 384 compared to DeBERTa-v3-small which features 6 layers of size 768.

Given these results, we fixed the sentence transformer to DeBERTa-v3-xsmall and evaluated the sum and weighted aggregations. Weighted aggregation yielded a notable improvement in classification performance and when compared to the original GraphNLI model (82.87% accuracy), this model achieved a meaningful gain of 0.27% accuracy, despite using a substantially smaller sentence transformer. To verify this performance, we applied weighted aggregation to all-MiniLM-L6-v2 and observed a similar relative improvement over its mean aggregation counterpart.

The model developed utilising the DeBERTa-v3-xsmall sentence transformer and the weighted aggregation method with a decay factor of 0.75 is considered our final model. This model achieved an accuracy of 83.14% and obtained a ROC AUC score of 0.9020, indicating strong class separation

ability, being only 0.098 off being a perfect classifier. Its confusion matrix illustrated class-specific performance differences. Out of 37000 actual attacking arguments, the model correctly predicted 31850 (true positives), and for the 27874 actual supporting arguments it correctly predicted 22001 (true positives). This disparity suggests further fine-tuning is warranted to improve recall for supporting arguments.

Models	Accuracy (%)	F1 Score (%)	Precision	Recall	ROC AUC
Bag of Words	67	67	67	67	0.75
Sentence BERT	78.86	78	78.43	78.86	0.855
MiniLM L6 V2 (Mean)	76.18	76.08	76.08	76.18	0.84
DistilRoBERTa Base (Mean)	78.48	78.46	78.47	78.48	0.86
DeBERTa-v3-small (Mean)	79.01	78.98	78.99	79.01	0.863
DeBERTa-v3-xsmall (Mean)	79.36	79.32	79.31	79.36	0.8706
DeBERTa-v3-xsmall (Sum)	79.48	79.45	79.44	79.48	0.8723
MiniLM L6 V2 (Weighted)	79.19	79.11	79.12	79.19	0.8667
DeBERTa-v3-xsmall (Weighted)	83.01	82.98	82.97	83.01	0.902

The complete performance metrics for all trained models are shown in the table above. The DeBERTa-v3-xsmall + weighted aggregation configuration consistently ranked highest, confirming it as the optimal model from our experiments.

The completion of the proposed ablation and sensitivity tests revealed the following properties:

1. **Walk length sensitivity (> or < four-walk)**

Walk lengths less than four reduce performance as they fail to capture enough context, removing informative ancestor nodes that help determine argument polarity. Larger walk lengths conversely provide minimal additional benefit while significantly increasing computational cost.

2. **Decay factor sensitivity (> or < 0.75)**

When decay factor is decreased (faster decay), potentially relevant mid-range ancestors are underweighted, while when it is increased (slower decay), distant ancestors become overweight adding noise and drift from the target argument.

3. **The removal of |u-v| within the Softmax head**

This substantially reduces accuracy as this feature was found to be the most informative concatenation component. It measures the semantic distance between the current argument's embedding (u) and its aggregated context vector (v).

4. **The increase of maximum sequence length beyond 256 tokens**

Increasing the variable max_seq_len beyond 256 did not yield meaningful gains and instead increased computational cost. This is because most Kialo arguments are concise, hence additional tokens typically capture padding or law-value context.

VI. CONCLUSION

This work presents a novel approach to argument move classification in online debates by using strategic graph walks, aggregation methods and sentence transformers. Our model successfully adapts the GraphNLI framework while improving both efficiency and performance for polarity prediction tasks.

Key improvements include:

Graph Context Integration: Weighted root-seeking walks significantly improved accuracy by using the weighted decay factor of 0.75. This makes the model more effective and confirms that proximal ancestors provide more relevant context than distant nodes.

Parameter Efficiency: Our model was trained with different parameters and sentence transformers and found that DeBERTa-v3-xsmall (22M parameters) delivers strong results while maintaining local deployment feasibility and enhancing efficiency with shorter training and testing time.

Architecture Trade-offs: MiniLM offers fastest inference while DeBERTa variants provide superior semantic understanding. However, DeBERTa-v3-xsmall (22M parameters) achieves the optimal balance of accuracy and efficiency, outperforming baselines on both metrics, suggesting significant potential for future architectural improvements with faster training and testing time.

Limitation and Future Work:

The current dataset is limited to binary classification (support/attack) on Kialo's highly moderated platform. The clean, structured nature of Kialo debates may not reflect the noisy, informal argumentation patterns found on platforms like Reddit or Twitter, where datasets would require manual annotation of polarity since they lack Kialo's structured labeling. Moreover, the fixed walk length may limit understanding of longer debate threads, though extending this might increase training time. **More sophisticated attention mechanisms for aggregation methods** could better identify specific contextual elements for improved results.

Additionally, training on different languages and specialized domains would provide valuable insights into argument styles from different regions of the world. In the future, the model could be trained with **DeBERTa-v3-Large** which contains more parameters and could yield optimal results if computational time is not constrained.

Finally, this work demonstrates the advancement from RoBERTa-base architectures to the efficient v3-xsmall model which can train faster and more accurately, establishing a new benchmark for parameter-efficient argument classification in structured debate environments.

REFERENCES

I. Agarwal, V. (2022) A graph-based natural language inference model for polarity ..., GraphNLI: A Graph-based Natural Language Inference Model for Polarity Prediction in Online Debates. Available at: https://arxiv.org/pdf/2202.08175 (Accessed: 28 July 2025).

II. Reimers, N. (2019) *Sentence embeddings using Siamese Bert-Networks*, *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. Available at: https://aclanthology.org/D19-1410.pdf (Accessed: 28 July 2025).

III. He, P., Liu, X., Gao, J. and Chen, W. (2021) DEBERTA: Decoding-Enhanced BERT with Disentangled Attention. *International Conference on Learning Representations*. Available at: https://openreview.net/forum?id=XPZIaotutsD (Accessed: 29 July 2025).

IV. Kialo Edu (n.d.) *Kialo Edu: The free tool for thoughtful, inclusive class discussion*. Available at: https://www.kialo-edu.com/ (Accessed: 29 July 2025).