# Enhancing Essay Argument Persuasiveness Prediction Using a RoBERTa-LSTM Hybrid Model

Fahad M. Alzaidee[1], Tommy Yuan[1], Peter Nightingale[1] and Khaled El Ebyary[1,*,†]

[1] *1University of York, Heslington, York YO10 5DD, United Kingdom*

**Abstract**

Over the past five decades, automated essay scoring has been a significant focus in both research and industry, capturing the interest of the NLP community due to its potential to provide valuable educational tools that save time for educators worldwide. The persuasiveness of arguments is a key aspect of argumentative essay quality. However, despite its significance, the persuasiveness of arguments has often been overlooked in research, with most studies still in their infancy. In this paper, we introduce several neural models aimed at enhancing the prediction of the persuasiveness score of an argument. Our proposed model improved the prediction accuracy compared to the approach suggested in [1].

**Keywords**

Natural Language Processing, Persuasiveness scoring, Argument evaluation

## 1. Introduction

In our globalizing world, learning English has become essential. Writing, a fundamental aspect of language learning, requires accurate assessment. Automated Essay Scoring (AES) offers a solution to the complex and time-consuming task of manual essay scoring. Even with standardised rubrics, manual scoring is often subjective and unreliable due to individual factors like mood and personality. AES provides an objective and efficient alternative, facilitating consistent evaluation of writing skills and supporting self-study through automated, unbiased, and instant feedback.

There are many types of essays, each serving different purposes and engaging readers in unique ways. This paper will focus on persuasive and argumentative essays, which are designed to convince the reader of a particular viewpoint through well-reasoned arguments and evidence. Argumentation, which can take various forms but generally involves presenting and defending a claim with supporting evidence or reasoning, is a critical skill for students to master. Effective argumentation not only strengthens academic performance but also equips students with essential communication skills for the real world. Automated Essay Scoring (AES) systems can play a significant role in this learning process by providing immediate, objective feedback, enabling students to refine their persuasive writing skills and develop stronger, more compelling arguments over time. Although previous studies (e.g., [1], [2]]) have explored automated essay scoring and feedback, further research is necessary to fully realize this vision.

In this study, we explore new NLP model architectures for automatically scoring argumentative essays based on their persuasiveness. We also augmented the standard dataset used in [1] by extracting arguments from each essay, paraphrasing them, and enriching them with supplementary information. These enhancements improve the dataset's overall quality and depth.

## 2. Related Work

Work on argument persuasiveness, closely related to our study, has been explored by several researchers. Persing and Ng [3] identify elements that weaken persuasiveness, while Stab and Gurevych [4] investigate the adequacy of argument support. Al Khatib et al. [5] annotate a news editorial corpus with

✉ fmfa501@york.ac.uk (F. M. Alzaidee); tommy.yuan@york.ac.uk (T. Yuan); peter.nightingale@york.ac.uk (P. Nightingale); khaled.elebyary@york.ac.uk (K. E. Ebyary)

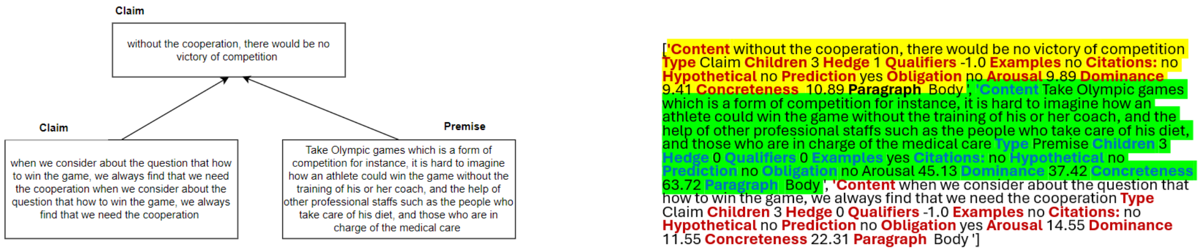
CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

detailed argumentative discourse units to examine persuasive strategies. Schaefer et al. [6] identify key factors that influence the persuasiveness of a text, including the usage patterns of argument components, the structure of the essay, the flow and sequence of argument types, as well as the impact of the essay prompt and the individual author's style. Additionally, Wachsmuth et al. [7] identify and annotate 15 dimensions—logical, rhetorical, and dialectical—relevant for automatically evaluating argument quality. Ke et al. [1] introduce an artificial neural network, bidirectional LSTM model with attention mechanisms, to score metrics like persuasiveness, specificity, and strength using a neural network on their annotated dataset of 102 essays [8]. Toledo et al. [9] publish a new dataset with arguments annotated for quality and compared arguments in pairs to determine the stronger one. They utilised a BERT language model to generate numerical representations for words in both arguments and then fine-tuned it for the classification and ranking tasks. In 2023, another study [10] used the PERSUADE dataset to predict persuasiveness ratings for discourse elements based on their type labels. Previous studies, such as [2], that evaluate the overall persuasiveness of an entire essay often provide generalized feedback, lacking the granularity needed to highlight specific areas for improvement. On the other hand, the study in [1] offers feedback on different traits impacting the persuasiveness of various essay sections but still demonstrates only modest performance.

## 3. Dataset

The aspect of persuasiveness in essays has been annotated in several available datasets, such as [11], [12], and [8]. However, datasets [11] and [12] lack the granularity required to train our model effectively. Therefore, we decided to use the dataset [8], which comprises 102 essays from the Annotated Essays corpus by Stab and Gurevych [13]. Each essay is annotated with an argument tree, mirrors the natural flow of argumentative essays, avoiding cycles and maintaining clarity. These trees typically have three to four levels, beginning with the Major Claim, followed by Claims and Premises that support or challenge their parent nodes. The dataset includes 1,459 components: 185 Major Claims, 567 Claims, and 707 Premises, each of which is assigned with various score metrics, including persuasiveness, which is the focus of our work. The Krippendorff's values for persuasiveness annotations (0.739 for Major Claims, 0.701 for Claims, and 0.552 for Premises) indicate that the dataset is well-suited for training our model. The dataset was split into training and testing sets, with the training set representing 80% of the essays.

To address the dataset's limited size, we identified all possible arguments in each set, generating 1,459 distinct arguments. We created two different sequences for each argument: one based on the order of appearance in its original essay and the other using postorder traversal. Inspired by [14], we enriched each component with lexical and structural features as illustrated in Figure **??**, increasing the maximum length from 58 to 85 words. Additionally, we paraphrased each argument component in the training set five times using a fine-tuned ChatGPT paraphraser on T5 (Text-to-Text Transfer Transformer). This resulted in four forms of input data: plain and enriched arguments using their order of appearance, and plain and enriched arguments using postorder traversal.


Claim
without the cooperation, there would be no victory of competition
Claim
when we consider about the question that how to win the game, we always find that we need the cooperation when we consider about the question that how to win the game, we always find that we need the cooperation
Premise
Take Olympic games which is a form of competition for instance, it is hard to imagine how an athlete could win the game without the training of his or her coach, and the help of other professional staffs such as the people who take care of his diet, and those who are in charge of the medical care

(a) Plain argument

['Content without the cooperation, there would be no victory of competition Type Claim Children 3 Hedge 1 Qualifiers -1.0 Examples no Citations: no Hypothetical no Prediction yes Obligation no Arousal 9.89 Dominance 9.41 Concreteness 10.89 Paragraph Body ', 'Content Take Olympic games which is a form of competition for instance, it is hard to imagine how an athlete could win the game without the training of his or her coach, and the help of other professional staffs such as the people who take care of his diet, and those who are in charge of the medical care Type Premise Children 3 Hedge 0 Qualifiers 0 Examples yes Citations: no Hypothetical no Prediction no Obligation no Arousal 45.13 Dominance 37.42 Concreteness 63.72 Paragraph Body ', 'Content when we consider about the question that how to win the game, we always find that we need the cooperation Type Claim Children 3 Hedge 0 Qualifiers -1.0 Examples no Citations: no Hypothetical no Prediction no Obligation yes Arousal 14.55 Dominance 11.55 Concreteness 22.31 Paragraph Body ']

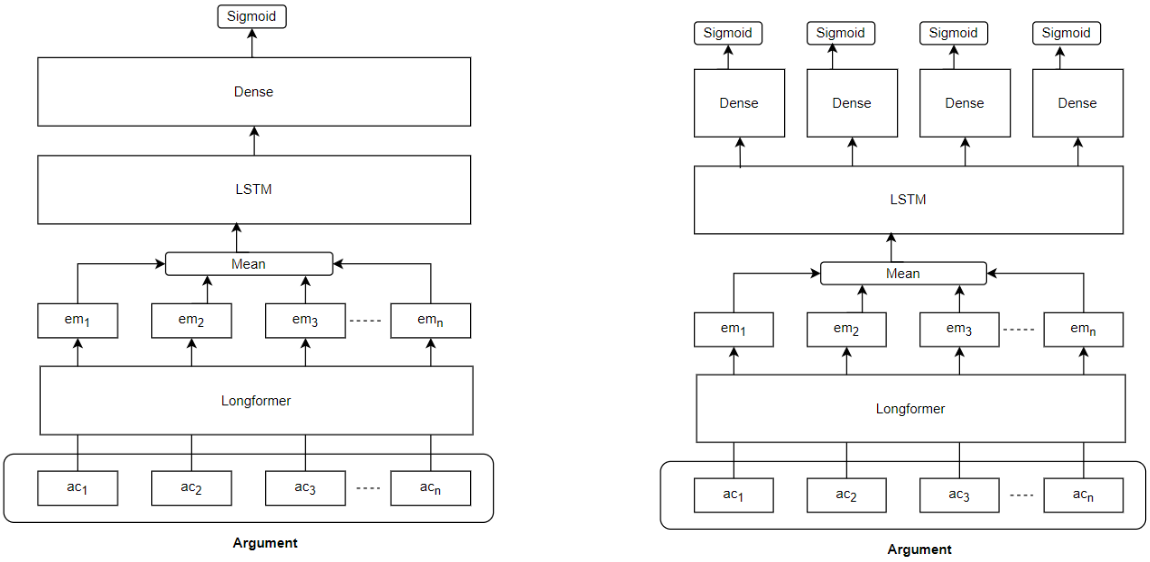
(b) Enriched argument

**Figure 1:** Enriching an argument

# 4. Proposed Methodology

In this study, we design four different neural models and compare their accuracy with two baseline models. Starting with baseline models, we fine-tuned the language model Longformer and the second model based on a Hierarchical BERT framework (HBM) [15] designed for classifying long documents with limited labelled data. In the HBM-based model, we identify the argument components in each argument and independently convert their tokens, which can be words or subwords, into their numerical vectors with the RoBERTa encoder. We use mean pooling to average these vectors, creating a single representation for each argument component. The new computed vectors are then input into the sentence-level Hierarchical BERT encoder, generating an intermediate representation of the entire argument. We adapted this model to predict persuasiveness scores as continuous values, by adding a sigmoid activation function after the linear layer, multiplying its output by 6, and rounding to the nearest integer.

We designed four neural models, illustrated in Figures **??** and **??**, combining a transformer model with an LSTM layer. RoBERTa and Longformer were used as embedding layers to generate a representation for each argument component in each argument, while the LSTM layer captured the dependency among the argument components. We refer to those models as LONG-LSTM, LONG-LSTM-TAG, ROB-LSTM, and ROB-LSTM-TAG. The term "TAG" in the model name indicates that the model uses a multi-output approach. In LONG-LSTM and LONG-LSTM-TAG, tokens for all argument components are embedded in a single pass using Longformer. Token embeddings are extracted and mean pooled to create a single representation for each component, which is then fed into an LSTM layer. In LONG-LSTM, the final hidden state is passed through a linear layer followed by a sigmoid activation function, producing a persuasiveness score between 0 and 6. The output is scaled by 6 and rounded to the nearest integer. MAE is then computed. In LONG-LSTM-TAG, each encoded argument component's hidden state is used to predict its persuasiveness score. ROB-LSTM and ROB-LSTM-TAG follow the same process except each argument component is encoded independently using RoBERTa.



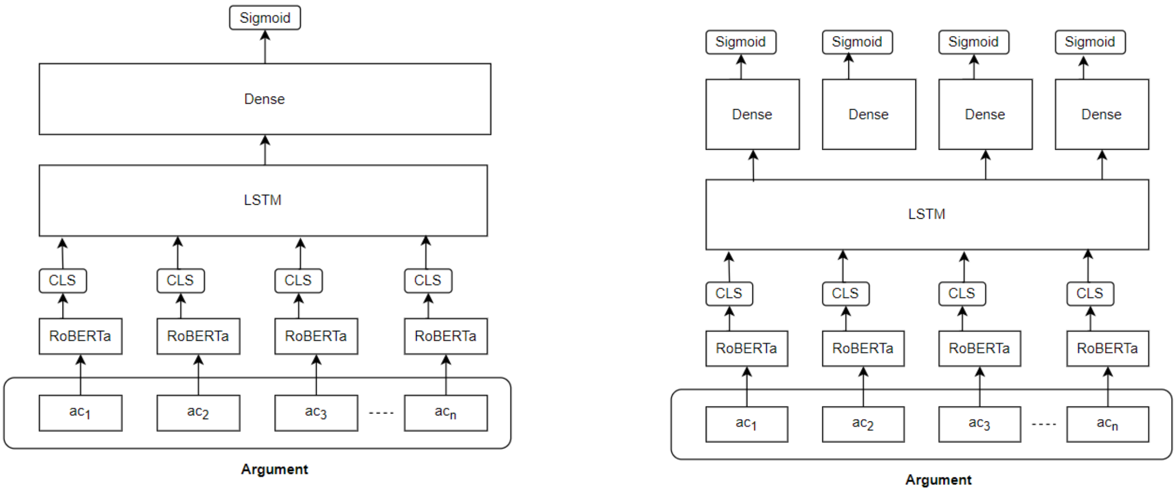
(a) LONG-LSTM

(b) LONG-LSTM-TAG

**Figure 2:** longformer-based models

# 5. Experiment Setup

We begin by randomly dividing our training set into five parts and perform five-fold cross-validation. In each experiment, four parts are used for training and one for development. After each iteration, we test



(a) ROB-LSTM

(b) ROB-LSTM-TAG

**Figure 3:** RoBERTa-based models

each model and compute the MAE and PC on the rounded predicted scores. The overall performance of the models is determined by averaging the metrics from all five iterations.

For training the HBM-based model, we used a learning rate of $2 \times 10^{-5}$, a dropout rate of 0.01, 50 epochs, a batch size of 4, and the Adam optimizer with a learning rate decay set to $1 \times 10^{-8}$. For other models, we used a learning rate of $1 \times 10^{-3}$, a dropout rate of 0.7, 50 epochs with a batch size of 4, and the Adam optimizer.

To evaluate the models' prediction accuracy, we used Mean Absolute Error (MAE) and Pearson's Correlation Coefficient (PC), computed after rounding the predicted scores. MAE measures the average distance between the predicted and actual scores. PC reflects the consistency and directionality of the predictions. We use the MAE instead of the Mean Squared Error (MSE) for its equal treatment of errors and reduced sensitivity to outliers. Additionally, it facilitates direct comparison with the models in [1].

## 6. Evaluating the Effectiveness of Models

Table 1 presents a summary of the evaluation experiments conducted on the test set. The leftmost column lists the various modeling approaches, while the top row identifies the different types of input data used. The HBM-based model shows the strongest correlations (PC) for plain arguments (0.309) and plain arguments (Postorder) (0.322). For rich arguments, the Fine-tuned Longformer achieves the highest correlation (0.702), while the ROB-LSTM-TAG model has the highest correlation (0.696) for rich arguments (Postorder).

The table clearly illustrates that incorporating rich features alongside argument content significantly improves model performance. Among the models, the ROB-LSTM stands out for its balanced performance, achieving a low MAE of 0.743 and a high PC of 0.691 in the 'Rich Argument (Postorder)' category. This suggests the potential effectiveness of leveraging hierarchical dependencies between argument components to predict persuasiveness. We also trained all models on the training set after paraphrasing each argument component, but this did not lead to improved results when tested on the test set.

The improvement in RoBERTA-based models is attributed to generating representations for each argument component independently, which reduces noise from other components in the same argument. In contrast, encoding the entire argument using the Longformer introduces more noise. Adding an LSTM layer further helps by separating the content of the argument component from its contextual dependency, thus reducing noise.

In comparison to the closest related work by [1], which was also trained on the same dataset and reported a MAE of 0.983 and a PC of 0.353, our ROB-LSTM model demonstrates a substantial

**Table 1**
Persuasiveness scores of all models on the test set

| Model | Plain Argument | | Plain Argument (Postorder) | | Rich Argument | | Rich Argument (Postorder) | |
|---|---|---|---|---|---|---|---|---|
| | MAE | PC | MAE | PC | MAE | PC | MAE | PC |
| HBM-based model | 2.478 | **0.309** | 2.532 | **0.322** | 2.078 | 0.441 | 2.025 | 0.443 |
| Fine-tuned Longformer | 1.297 | 0.284 | 1.351 | 0.238 | 0.919 | **0.702** | 0.932 | 0.664 |
| LONG-LSTM | 1.419 | 0.125 | 1.257 | 0.320 | 1.000 | 0.566 | 1.041 | 0.513 |
| LONG-LSTM-TAG | 1.311 | 0.124 | 1.311 | 0.159 | 0.835 | 0.612 | 0.880 | 0.598 |
| ROB-LSTM | 1.230 | 0.251 | 1.203 | 0.223 | **0.757** | 0.685 | **0.743** | **0.691** |
| ROB-LSTM-TAG | **1.137** | 0.213 | **1.146** | 0.198 | 0.790 | 0.604 | 0.747 | 0.646 |

improvement, achieving a $24.4\%$ reduction in MAE and an $95.8\%$ increase in PC.

## 7. Conclusion and Future Work

In this preliminary study, we explored different models to predict the persuasiveness score of arguments with varying complexity and structures. The RoBERTa-LSTM model demonstrated a balanced performance where it achieved a low MAE and a relatively high PC. The addition of rich features and the consideration of hierarchical order relations highlighted the potential benefits of these factors in improving persuasiveness prediction.

A significant challenge we face is incorporating the types of relationships between arguments within an essay. We aim to understand how the persuasiveness of lower-level arguments in the argument tree affects the overall persuasiveness of related higher-level arguments. This understanding is crucial for developing a feedback component in our system that effectively leverages these relationships.

## References

[1] Z. Ke, W. Carlile, N. Gurrapadi, V. Ng, Learning to give feedback: Modeling attributes affecting argument persuasiveness in student essays., in: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, 2018, pp. 4130–4136.

[2] T. Wambsganss, C. Niklaus, M. Cetto, M. Söllner, S. Handschuh, J. M. Leimeister, Al: An adaptive learning support system for argumentation skills, in: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 2020, pp. 1–14.

[3] I. Persing, V. Ng, Why can't you convince me? modelling weaknesses in unpersuasive arguments, in: Proceedings of the 26th International Joint Conference on Artificial Intelligence, 2017, pp. 4082–4088.

[4] C. Stab, I. Gurevych, Recognizing insufficiently supported arguments in argumentative essays, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, 2017. URL: https://api.semanticscholar.org/CorpusID:6801402.

[5] K. Al Khatib, H. Wachsmuth, J. Kiesel, M. Hagen, B. Stein, A news editorial corpus for mining argumentation strategies, in: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, 2016, pp. 3433–3443.

[6] R. Schaefer, R. Knaebel, M. Stede, Towards fine-grained argumentation strategy analysis in persuasive essays, in: M. Alshomary, C.-C. Chen, S. Muresan, J. Park, J. Romberg (Eds.), Proceedings of the 10th Workshop on Argument Mining, Association for Computational Linguistics, Singapore, 2023, pp. 76–88. URL: https://aclanthology.org/2023.argmining-1.8. doi:10.18653/v1/2023.argmining-1.8.

[7] H. Wachsmuth, N. Naderi, Y. Hou, Y. Bilu, V. Prabhakaran, T. Alberdingk Thijm, G. Hirst, B. Stein, Computational argumentation quality assessment in natural language, in: Proceedings of the 15th

Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, 2017, pp. 176–187.

[8] W. Carlile, N. Gurrapadi, Z. Ke, V. Ng, Give me more feedback: Annotating argument persuasiveness and related attributes in student essays, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 621–631.

[9] A. Toledo, S. Gretz, E. Cohen-Karlik, R. Friedman, E. Venezian, D. Lahav, M. Jacovi, R. Aharonov, N. Slonim, Automatic argument quality assessment–new datasets and methods, arXiv preprint arXiv:1909.01007 (2019).

[10] Y. Hicke, T. Tian, K. Jha, C. H. Kim, Automated essay scoring in argumentative writing: Deberteachingassistant, arXiv preprint arXiv:2307.04276 (2023).

[11] C. Stab, I. Gurevych, Recognizing insufficiently supported arguments in argumentative essays, in: M. Lapata, P. Blunsom, A. Koller (Eds.), Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, Association for Computational Linguistics, Valencia, Spain, 2017, pp. 980–990. URL: https://aclanthology.org/E17-1092.

[12] S. Li, V. Ng, ICLE++: Modeling fine-grained traits for holistic essay scoring, in: K. Duh, H. Gomez, S. Bethard (Eds.), Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 8465–8486. URL: https://aclanthology.org/2024.naacl-long.468. doi:10.18653/v1/2024.naacl-long.468.

[13] C. Stab, I. Gurevych, Annotating argument components and relations in persuasive essays, in: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, 2014, pp. 1501–1510.

[14] U. Mushtaq, J. Cabessa, Argument classification with bert plus contextual, structural and syntactic features as text, in: International Conference on Neural Information Processing, Springer, 2022, pp. 622–633.

[15] J. Lu, M. Henchion, I. Bacher, B. Mac Namee, A sentence-level hierarchical bert model for document classification with limited labelled data, in: Discovery Science: 24th International Conference, DS 2021, Halifax, NS, Canada, October 11–13, 2021, Proceedings, Springer, 2021, pp. 231–241.