Developing a Transformer-Based Automated Essay Scoring System with Explainability Considerations

Oscar Dos Santos Nunes

March 2025 - September 2025

Contents

Chapter 1

Introduction

1.1 Background and Context

1.2 Background and Context

Automated Essay Scoring (AES) systems have been in development for over fifty years, evolving from statistical and rule-based approaches into increasingly sophisticated machine learning and deep learning technologies. Early developments such as Project Essay Grade (PEG) (Page 1966) and e-rater (Attali and Burstein 2006) relied on handcrafted features, including high-level syntactic and lexical features. While these methods demonstrated some success and saw real-world application, they often struggled to capture the semantic and contextual depth found in student writing (Shermis and Burstein 2013).

With the growth of more data-focused approaches in natural language processing (NLP), AES has increasingly adopted machine learning techniques. Traditional models such as support vector machines (SVM), random forests, and ridge regression were trained on handcrafted features to predict human-marked essay scores. However, these models were still limited by their reliance on manually created inputs, which creates constraints with scalability and generalisation across writing prompts and student population groups.

The growth of deep learning, particularly with transformer-based models, has significantly advanced AES research. Pre-trained language models such as BERT (Devlin et al. 2019) and RoBERTa (Liu et al. 2019) have demonstrated impressive results across various NLP tasks, thanks to their ability to model contextual relationships within text. When compared to more typical machine learning techniques, studies using BERT for AES have shown substantial gains in accuracy, especially when using benchmark datasets such as the ASAP (Automated Student Assessment Prize) corpus (Taghipour and Ng 2016).

Despite these advancements, a number of obstacles still exist. Deep learning models frequently act as "black boxes", providing little to no room for interpretation. This is a major drawback for AES, as user trust and educational integration rely on the reasoning

behind a score. In recent years, there has also been growing concern about algorithmic bias, particularly when scoring responses from under-represented student groups or non-native English speakers (Blodgett et al. 2020). These issues raise questions about fairness, generalisability, and the ethical use of AES in high-stakes contexts.

Furthermore, much of current AES work focuses on holistic scoring without exploring the formative aspects of feedback, rubric-based scoring, or the integration of multi-modal inputs. Expandability methods such as SHAP (Lundberg and Lee 2017) or attention-based visualisations have been actively researched as a means of improving model transparency. Hybrid systems that combine neural and symbolic components have also shown promise in enhancing interpretability and control.

This project builds on these advancements by developing an AES model using deep learning, evaluating its performance and fairness, and investigating potential pathways for expansion. Through comparing results against traditional baselines and considering explainability and ethical implications, this work aims to contribute to research seeking to make AES systems more robust, transparent, and educationally useful.

1.3 Problem Statement

1.4 Aims and Objectives

1.5 Research Questions

1.6 Significance of the Study

1.7 Dissertation Structure Overview

Chapter 2

Literature Review

2.1 Historical Evolution of Automated Essay Scoring

2.2 Rule-Based and Traditional ML Methods

2.3 Deep Learning and Transformer-Based Models

2.4 Datasets for AES

2.5 Evaluation Metrics in AES

2.6 Challenges in AES

2.7 Gaps in Existing Research

Chapter 3

Methodology

3.1 Research Design and Approach

3.2 Data Collection and Preprocessing

3.2.1 Dataset Selection and Description

3.2.2 Tokenisation and Cleaning

3.2.3 Label Normalisation

3.3 Model Architecture and Implementation

3.3.1 Baseline Models

3.3.2 Transformer Fine-Tuning

3.3.3 Training Configuration

3.4 Evaluation Strategy

3.4.1 Metrics Used

3.4.2 Cross-Validation Setup

3.5 Tools and Libraries

3.6 Ethical Considerations

Chapter 4

Experiments and Results

4.1 Baseline Performance

4.2 Transformer Performance

4.3 Error Analysis

4.4 Model Explainability

4.5 Fairness and Bias Analysis

4.6 Comparison with Human Raters

Chapter 5

Discussion

5.1 Interpretation of Results

5.2 Contributions to the Field

5.3 Limitations

5.4 Implications for Educational Technology

Chapter 6

Conclusion

6.1 Summary of Findings

6.2 Answers to Research Questions

6.3 Future Work

Bibliography

Attali, Yigal and Jill Burstein (2006). "Automated Essay Scoring With e-rater® V.2". In: *The Journal of Technology, Learning and Assessment* 4.3, pp. 1–21. URL: https://ejournals.bc.edu/index.php/jtla/article/view/1650.

Blodgett, Su Lin et al. (2020). "Language (Technology) is Power: A Critical Survey of "Bias" in NLP". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5454–5476. URL: https://aclanthology.org/2020.acl-main.485/.

Devlin, Jacob et al. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *arXiv preprint arXiv:1810.04805*. URL: https://arxiv.org/abs/1810.04805.

Liu, Yinhan et al. (2019). "RoBERTa: A Robustly Optimized BERT Pretraining Approach". In: *arXiv preprint arXiv:1907.11692*. URL: https://arxiv.org/abs/1907.11692.

Lundberg, Scott M. and Su-In Lee (2017). "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems* 30, pp. 4765–4774. URL: https://arxiv.org/abs/1705.07874.

Page, Ellis Batten (1966). "The Imminence of Grading Essays by Computer". In: *Phi Delta Kappan* 47.5, pp. 238–243. URL: https://wac.colostate.edu/docs/books/usu/machine/chapter17.pdf.

Shermis, Mark D. and Jill Burstein (2013). *Handbook of Automated Essay Evaluation: Current Applications and New Directions*. Routledge. URL: https://www.routledge.com/Handbook-of-Automated-Essay-Evaluation-Current-Applications-and-New-Directions/Shermis-Burstein/p/book/9780415810968.

Taghipour, Kaveh and Hwee Tou Ng (2016). "A Neural Approach to Automated Essay Scoring". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1882–1891. URL: https://aclanthology.org/D16-1193/.

Appendix A

Code Snippets

Appendix B

Gantt Chart / Timeline

Appendix C

Additional Tables or Figures

Appendix D

Ethics Approval Form