

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

Oscar Dos Santos Nunes

September 2025

# Contents

<b>1</b>	<b>Introduction</b>	<b>3</b>
1.1	Background and Context . . . . .	3
1.2	Problem Statement . . . . .	3
1.3	Aims and Objectives . . . . .	3
1.4	Research Questions . . . . .	3
1.5	Significance of the Study . . . . .	3
1.6	Dissertation Structure Overview . . . . .	3
<b>2</b>	<b>Literature Review</b>	<b>4</b>
2.1	Historical Evolution of Automated Essay Scoring . . . . .	4
2.2	Rule-Based and Traditional ML Methods . . . . .	4
2.3	Deep Learning and Transformer-Based Models . . . . .	4
2.4	Datasets for AES . . . . .	4
2.5	Evaluation Metrics in AES . . . . .	4
2.6	Challenges in AES . . . . .	4
2.7	Gaps in Existing Research . . . . .	4
<b>3</b>	<b>Methodology</b>	<b>5</b>
3.1	Research Design and Approach . . . . .	5
3.2	Data Collection and Preprocessing . . . . .	5
3.2.1	Dataset Selection and Description . . . . .	5
3.2.2	Tokenisation and Cleaning . . . . .	5
3.2.3	Label Normalisation . . . . .	5
3.3	Model Architecture and Implementation . . . . .	5
3.3.1	Baseline Models . . . . .	5
3.3.2	Transformer Fine-Tuning . . . . .	5
3.3.3	Training Configuration . . . . .	5
3.4	Evaluation Strategy . . . . .	5
3.4.1	Metrics Used . . . . .	5
3.4.2	Cross-Validation Setup . . . . .	5
3.5	Tools and Libraries . . . . .	5

---

3.6	Ethical Considerations . . . . .	5
<b>4</b>	<b>Experiments and Results</b>	<b>6</b>
4.1	Baseline Performance . . . . .	6
4.2	Transformer Performance . . . . .	6
4.3	Error Analysis . . . . .	6
4.4	Model Explainability . . . . .	6
4.5	Fairness and Bias Analysis . . . . .	6
4.6	Comparison with Human Raters . . . . .	6
<b>5</b>	<b>Discussion</b>	<b>7</b>
5.1	Interpretation of Results . . . . .	7
5.2	Contributions to the Field . . . . .	7
5.3	Limitations . . . . .	7
5.4	Implications for Educational Technology . . . . .	7
<b>6</b>	<b>Conclusion</b>	<b>8</b>
6.1	Summary of Findings . . . . .	8
6.2	Answers to Research Questions . . . . .	8
6.3	Future Work . . . . .	8
	<b>References</b>	<b>9</b>
<b>A</b>	<b>Code Snippets</b>	<b>10</b>
<b>B</b>	<b>Gantt Chart / Timeline</b>	<b>11</b>
<b>C</b>	<b>Additional Tables or Figures</b>	<b>12</b>
<b>D</b>	<b>Ethics Approval Form</b>	<b>13</b>

# Chapter 1

## Introduction

1.1 Background and Context

1.2 Problem Statement

1.3 Aims and Objectives

1.4 Research Questions

1.5 Significance of the Study

1.6 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

# References

# Appendix A

## Code Snippets

# Appendix B

## Gantt Chart / Timeline

# Appendix C

## Additional Tables or Figures

## Appendix D

### Ethics Approval Form