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

Oscar Dos Satos Nunes

MSc In Data Science
The University of Bath
2024 - 2025

⟨Dissertation Title⟩

Submitted by: ⟨Student Name⟩

Copyright

Declaration

This dissertation is submitted to the University of Bath in accordance with the requirements of the degree of Master of Science in the Department of Computer Science. No portion of the work in this dissertation has been submitted in support of an application for any other degree or qualification of this or any other university or institution of learning. Except where specifically acknowledged, it is the work of the author.

Abstract

⟨The abstract should appear here. An abstract is a short paragraph describing the aims of the project, what was achieved and what contributions it has made.⟩

Contents

List of Figures

List of Tables

Acknowledgements

Add any acknowledgements here.

List of Abbreviations

AAE	Argument Annotated Essays corpus
AES	Automated Essay Scoring
ASAP	Automated Student Assessment Prize corpus
BERT	Bidirectional Encoder Representations from Transformers
ELL / ESL	English Language Learner / English as a Second Language
ETS	Educational Testing Service (developers of *e-rater*)
FCE	First Certificate in English learner-essay corpus
HAN	Hierarchical Attention Network
IAM	Institut für Informatik A und B Handwriting Database
MAE	Mean Absolute Error
MOOC	Massive Open Online Course
MSE	Mean Squared Error
NLP	Natural-Language Processing
PEG	Project Essay Grade
QWK	Quadratic Weighted Kappa
RNN	Recurrent Neural Network
RMSE	Root Mean Squared Error
RoBERTa	Robustly Optimised BERT Approach
SHAP	SHapley Additive exPlanations
SVM	Support Vector Machine
TOEFL11	Non-native English Essay Corpus with 11 L1 backgrounds

Chapter 1

Introduction

Modern AES research sits at the intersection of educational measurement, natural-language processing (NLP), and responsible AI. Accordingly, this dissertation sets out to (i) examine how the field evolved from rule-based scoring to transformer models, (ii) develop and evaluate a RoBERTa-based scorer, and (iii) examine its transparency, fairness, and practical viability in real-world learning situations.

1.1 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 (Zhang and Litman, 2020). 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, 2013b).

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 (Shermis and Burstein, 2013b). 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 (Zhang and Litman, 2020).

The growth of deep learning, particularly with transformer-based models, has significantly advanced AES research (Vaswani et al., 2017). 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 (Lundberg and Lee, 2017). 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.

1.2 Problem Statement

The rapid growth of digital education and online learning platforms has brought a fresh interest in an efficient, scalable and consistent method for assessing students' work. Essay writing remains one of the vital methods for evaluating students' understanding of content as well as their ability to communicate ideas and think critically. However, this method requires significant time and resources from educators when marked manually. Human grading is not only more labour-intensive but it is also prone to consistency issues due to marker variability, fatigue and personal bias (Page and Petersen, 2003). In large-scale settings, the demand for a scalable assessment has led to the exploration of AES systems which are capable of mimicking human judgement in evaluating written answers (Burstein, Chodorow and Leacock, 2003).

Despite the decades of research and development within the field (ranging from early rule-based systems to more modern neural linguistic approaches), there are still significant limitations(Dikli, 2006). Traditional AES systems such as e-rater (Attali and Burstein, 2006) and Project Essay Grade (PEG) (Page and Petersen, 2003) relied on shallow linguistic features (e.g. word count, spelling, grammar) that do not capture deeper semantic and argument quality (Dikli, 2006). More recent advancements with transformer-based models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) offer promising alternatives through leveraging contextual embeddings and attention mechanisms to better model the meaning and coherence of a given text (Devlin et al., 2019).

However, these models still struggle with generalisation across writing prompts, transparency within their decision-making, and potential still hold bias against certain writing styles or demographic groups Beigman Klebanov et al. (2021b).

The core problem of this project seeks to address, is how to design and evaluate an AES model that offers both high predictive accuracy and with interpretability, while remaining robust across different prompts. In particular this project will examine:

- The efficiency of fine-tuned transformers models for essay scoring tasks;
- The generalisability of models across prompt types and writing levels;

- The trade off between model complexity, performance and interoperability;

This problem is significant both in an academia and practical context. From the perspective of research, it advances the state-of-the-art in educational NLP and model explainability. From a practical standpoint, it offers potential improvements in the fairness and scalability of the assessment tools used in modern educational institutions, especially focusing on the needs of large-scale online learning platforms such as Coursera and edX.

1.3 Aims and Objectives

The aim of this project is to develop a transformer-based AES system of which achieves high predictive performance while improving upon current work in interpretability and fairness. Through the use of state-of-the-art deep learning models such as BERT and RoBERTa, this work hopes to contribute towards more accurate, trustworthy and scalable alternatives to the more traditional essay marking systems (Devlin et al., 2019).

The primary objectives of this project are as follows:

1. **Review and evaluate existing work:** Conduct a comprehensive review of traditional and modern AES systems, focusing on their performance, limitations and explainability.
2. **Data selection and Preprocessing:** Identify appropriate datasets for training and evaluation. Preprocess essays through tokenization, cleaning and normalize score variations to enable effective modelling at later stages.
3. **Model development and comparisons:** Implement and fine-tune a transformer-based model for essay scoring. Develop a performance baseline off traditional techniques and previous AES systems for comparison.
4. **Performance and robustness evaluation:** Asses the accuracy of each models on standard performance metrics such as Quadratic Weighted Kappa (QWK) and Mean Squared Error. Evaluation the models across different prompts and writing levels.
5. **Explainability and transparency:** Apply interpretability techniques such as SHAP or attention-based visualisation to provide insights into model behaviours and decision making processes.
6. **Bias and fairness analysis:** Systematically evaluate potential demographic or linguistic biases within the models predictions. Explore the impact of these biases and evaluate potential solutions.
7. **Explore extensibility:** Discuss and develop potential pathways for extending model performances. Such as feedback generation, rubric-based scoring or the adaption of multi model inputs (Audio and handwritten answers).

1.4 Research Questions

This research is guided by the following key questions, which collectively shape the development, evaluation, and potential impact of the proposed Automated Essay Scoring (AES) system:

1. **To what extent can transformer-based models outperform traditional machine learning approaches in essay scoring tasks?** This question examines the predictive

performance of deep learning models such as BERT and RoBERTa in comparison to classical baselines like ridge regression or support vector machines.

2. **How well do transformer-based AES systems generalise across different essay prompts and writing levels?** Here, the focus is on evaluating model robustness and adaptability across diverse input types, including prompt-specific and cross-prompt scenarios.
3. **What trade-offs exist between predictive accuracy, model complexity, and interpretability in AES?** This question investigates whether performance gains from complex neural models justify potential losses in transparency and ease of understanding.
4. **Can model explainability techniques make AES systems more interpretable and trustworthy in educational settings?** This explores the practical value of applying tools such as SHAP or attention visualisations to clarify how model predictions are generated.
5. **What forms of algorithmic bias are present in AES outputs, and how can they be detected and mitigated?** This question addresses the ethical dimension of AES systems, particularly their fairness across student demographics and linguistic backgrounds.
6. **Is it feasible to extend the AES system with additional features, such as feedback generation or support for handwritten inputs?** This final question considers the future potential of AES models to go beyond scoring, towards becoming comprehensive tools for formative assessment.

1.5 Significance of the Study

AES systems have an ever-increasing importance within education, particularly with the rise of online learning environments such as MOOCs (Massive Open Online Courses), remote assessments, and digital feedback systems. Traditional human methods are heavily time-consuming, expensive, and prone to inconsistency. As a result, the development of a scalable, fair, and explainable AES system is a growing concern for educational tech.

This study contributes to the field through advancing the application of transformer-based deep learning models to the domain of essay evaluation. While prior research has demonstrated the performance benefits of these models, relatively little work has been examined for its transparency, fairness, and practical applicability. This project directly addresses that gap by incorporating explainability and fairness analysis into the model's development.

Furthermore, the study explores the trade-off between a model's accuracy and its interpretability, a critical consideration for its potential use within education. Through this comparison, this research will aid with guiding future work to mitigate biases and provide additional baselines for performance across a range of demographics.

Practically, the findings may support the development of a more robust essay evaluation tool for institutions seeking a new approach to assessment marking methods. Academically, it aids with the broader research into explainable AI and bias mitigation within educational NLP research.

1.6 Dissertation Structure Overview

This dissertation is organised as follows.

- Chapter 2 reviews prior work on rule-based, statistical and transformer-based AES systems, plus current explainability and fairness toolkits.
- Chapter 3 describes the datasets and preprocessing pipeline adopted in this study.
- Chapter 4 (Methodology) details baseline models, the fine-tuning strategy for RoBERTa, and the evaluation protocol—including fairness audits and SHAP analyses.
- Chapter 5 presents the experimental results, error analysis, and interpretability findings.
- Chapter 6 discusses implications, limitations, and avenues for extension to multimodal and multilingual scoring.
- Chapter 7 closes with key takeaways and recommendations for educators and researches.

Chapter 2

Literature Review

2.1 Historical Evolution of Automated Essay Scoring

The first notable milestone within AES was Ellis Page's development of Project Essay Grade (PEG) in the 1960s, pioneering the idea of using computers to assess writing quality. PEG worked through the identification of measurable surface features (such as word count, spelling, and grammar) as measures of writing quality, then fitting a regression model in order to predict human-assigned scores. Somewhat surprisingly, with such simple processes, Page reported correlations of around 0.7 between the PEG's scores and that of human raters. While this early work was far lacking in depth compared to that of today's, it demonstrated the "imminence of grading essays by computer" (Page, 1966) and generated excitement about the opportunities of algorithms judging complex writing.

After a significant break in research, in the 1990s many more AES systems began being developed. This was largely due to the significant advancements made within computing power and NLP research. One of the more notable advancements was e-rater, developed by the Educational Testing Service (ETS), which was first introduced in 1999, utilising a large set of linguistic features in order to mimic the scoring of a human marker (Attali and Burstein, 2006). Around the same time, other paradigms emerged: the Intelligent Essay Assessor (IEA) applied latent semantic reference tests (LSA) to assess content's relevance by comparing the semantic similarity of an essay to a corpus of reference texts. This content-focused approach complemented the earlier surface-feature methods (Landauer, Laham and Foltz, 2003), demonstrating the shift towards capturing more of an essay's substance and meaning in the automated scoring process.

In the early 2000s, research into AES increasingly included machine learning algorithms capable of leveraging both surface and content features. The more traditional statistical models were augmented and/or replaced by techniques such as support vector machines (SVM) and Bayesian classifiers, which were trained on large datasets of essays to optimise their scoring accuracy (Shermis and Burstein, 2013b). A key event for research was the **Hewlett Foundation's Automated Student Assessment Prize (ASAP)** competition in 2012 (Kaggle, 2012), this released a corpus of 13,000 scored essays across eight prompts and accelerated progress and benchmarking of what was possible with handcrafted features. In the mid-2010s a new breakthrough was found with the development of deep learning. Neural network-based AES models (beginning with Taghipour and Ng's 2016 RNN model) learned essay representations

from the raw text, rather than relying on hand-created features. These models achieved scoring accuracies that were comparable to human raters, often outperforming the previously developed algorithm (Taghipour and Ng, 2016). This rise of deep learning also renewed work on cross-domain scoring (training the model on one prompt but assessing work on another prompt), which had largely been abandoned in the early 2000s due to the lack of generalisability of the models (Beigman Klebanov et al., 2021a).

Alongside the improvements in accuracy, more recent work has brought attention towards AES systems' fairness, validity, and explainability. It is increasingly recognised that AES systems must not only just be reliable but also fair and transparent with its actions. Recent work has analysed for algorithmic favouring or penalising of certain groups. For example, studies have checked for biases in scores related to demographics (e.g. native language or gender), finding that in some cases performance can be significantly skewed (Blodgett et al., 2020). In parallel, research also stresses the importance of making the "black box" of deep learning models more interpretable. The latest systems attempt to provide feedback on the writing tasks, such as breaking the score down into a series of sub-scores looking at organisation, grammar, content and style. This push for explainability is driven by the need for trust and accountability within an educational context, where students and educators require an understanding of how the system arrived at its scores (Lundberg and Lee, 2017).

In summary, over the last 50 years, AES has developed from simple rule-based systems to complex deep learning models. The field has shifted from a focus on surface features to a more nuanced understanding of content and context, with an ever growing focus on fairness and explainability. This evolution reflects the broader trends in NLP and machine learning, as well as the increasing importance of ethical considerations in AI applications.

2.2 Rule-Based and Traditional ML Methods

Prior to the adoption of deep learning, AES systems were primarily developed using rule-based techniques and classic ML algorithms (Shermis and Burstein, 2013b). These systems relied on predefined features, crafted by expects, in order to model aspects of writing quality such as grammar, coherence, structure and vocabulary.

Early AES systems, such as PEG and e-rater, used deterministic formulae for scoring of which combine metrics like word count, average sentence length, character diversity and error frequency. These system offered interpretable outputs and performed well within their constrained environments. However, they required significant manual effort in order to design and maintain. Their reliance on shallow indicators led to them being brittle and often poorly generalised across different prompts, genres and writer population groups.

As computational methods improved, statistic based ML models replaced the hard-coded scoring rules. Algorithms such as linear regression, decision trees, random forests and SVMs trained on feature vectors extracted from essays. These features commonly included syntactic parse statistics, discourse structure elements and genre specific indicators aligning with the human rubrics (Taghipour and Ng, 2016).

Although these models improved predictive flexibility and allowed for prompt adaption, they still suffered from key limitations. Notably, the require a significant amount of feature engineering, highly sensitive to feature selection quality and lacked the ability to capture semantic meaning. As a result, they often failed to generalise to new prompts and writing styles, as well as

struggling with deeper linguistic aspects such as coherence, argument structure and tone.

It was these limitations that ultimately created the shift toward neural network approaches, of which aimed to learn high-level representations directly from text. The introduction of deep learning marked as a major turning point in AES development, reducing reliance on manual features and offer improved generalisability, scalability and scoring accuracy.

2.3 Deep Learning and Transformer-Based Models

The introduction of deep learning marked a key turning points for AES systems, allowing for models to learn features directly from raw text over the previous reliance on hand-crafted inputs. Early neural networks such as Taghipour and Ng's(2016) RNN-rule based approach achieved near-human scoring through modelling sequential dependencies in the texts. Hierarchical and attention-based networks followed this, which captured higher-level essay structures and enabled better interpretability (Yang et al., 2016). These architectures offered improved generalisation compared to traditional methods, however the performance gains where still limited by their sequential processes.

Transformer based models such as BERT and RoBERTa brought a significant boost to AES system capabilities. The models utilise self attention mechanisms to encode complex relationships within the text, allowing for an understanding of semantics, coherence and broad patters more effectively. Fine-tuning transformers on AES datasets such as ASAP has consistently outperformed prior approaches across both accuracy and generalisability benchmarks (Taghipour and Ng, 2016).

However, while transformers has provided strong predictive performance, they introduce many challenges with regards to explainability and fairness. These models act as "black boxes" with very limited transparency as to how a decision is made(Lundberg and Lee, 2017). This is highly problematic within the setting of education as stakeholders will require an insight into scoring rationales. Furthermore, models trained on specific prompts may not generalise well, and there are many concerns about potential algorithmic-biases that may favour or penalise certain groups of students.

Regardless of these drawbacks, transformer based models are currently the state-of-the-art within AES systems. The ever ongoing research into explainable AI and cross-prompt generalisation is begging to address these shortcomings, suggesting that more robust and fair systems are on the horizon.

2.3.1 Transformer-Based Models and Contextual Embeddings

The shift form recurrent networks to self-attention based architectures has reshaped the field of AES research. Building on the encoder introduced by Vaswani et al. (Vaswani et al., 2017), pre-trained language models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) encode deep contextual embeddings without the requirements of hand crafted features. Fine tuning these on essay corpora yields QWK gains of up to 0.10 over ridge-regression based model baselines (Taghipour and Ng, 2016).

Prompt-agnostic transformers approaches have also retained around 90% of their performance after label normalisation, suggesting stronger capabilities of generalisation in comparisons to RNNs or the Hierarchical Attention Networks (HANs) Transformer encoders are therefore

the standardised backbone of modern AES; this project adopts a RoBERTa-base variant for developing prompt-agnostic scoring.

2.3.2 Explainability Technologies

Interpretability is vital for development in an educational environment. SHAP (Lundberg and Lee, 2017) offers model-agnostic local explanations by identifying token-level importance through Shapley values. Attention roll-out and Integrated Gradients provide a less computational complex alternatives, however not always as reliable. Counterfactual perturbations are used to aid in the verification of models attesting to rubric-relevant markers over the prompt-specific features.

Among the available techniques SHAP provided consistent, reliable and human-readable rationale, aligning with the projects goals of improving transparency.

2.3.3 Fairness Auditing Toolkits and Metrics

Bias analysis in AES typically refers to QWK gaps, equal-opportunity difference and calibration errors. Open-source libraries such as aequitas and fairlearn automate these metrics and provide visual diagnostics (Litman, Baffour and Crossley, 2024).

Recent studies have shown significant score discrepancies between native and non-native English writers, as well as across different genders(Blodgett et al., 2020). Mitigation strategies include data re-balancing, adversarial training and post-hoc calibration.

These toolkits and metrics give a strong method for this projects work on bias analysis, giving us the resources to systematically detect model biases and work to mitigate these disparities in predicted scores.

2.4 Datasets for AES

AES research has been fuelled by the availability of large, annotated essay datasets. Commonly the Automated Student Assessment Prize (ASAP) corpus is used, released by the Hewlett Foundation in 2012. Consisting of almost 13,000 essays across eight prompts, each with prompt-specific scoring rubrics, ASAP has become the benchmark for AES systems (Kaggle, 2012). It enabled both prompt-specific and prompt-agnostic evaluations and remains as a standard for comparative performative analysis.

An alternative to ASAP is the TOEFL11 corpus, containing essays written by non-native English speakers (Blanchard et al., 2013). This data set is often used for evaluating a systems ability to handle second-language writing. The FCE dataset by Cambridge learner corpus also supports trait-based scoring with fine-grained annotations on vocabulary, cohesion and grammar (Yannakoudakis, Briscoe and Medlock, 2011). These datasets combined with others aid in research into both holistic and analytic scoring, aiding with the expansion of AES to a broader range of writing tasks.

More recently, datasets such as ASAP++ has added multi-trait labels, enabling models to be move beyond holistic scoring towards more detailed feedback generation. Along side this there are also corpus tailored for specific writing dimensions such as Argument Annotated Essays (AAE) which evaluate the persuasiveness and structure of arguments within text. Resources

such as these have supported the development of systems capable of evaluating across different genres and to comprehend a wider range of instruction goals.

Despite these advancements, much of the AES datasets are still limited to English and typed responses, with very few multi-lingual datasets or multi-modal datasets available. This significantly restrict our models ability to generalize across languages and formats. Expanding this landscape of datasets to be more inclusive remains a key step in developing a Generic AES system.

2.5 Evaluation Metrics in AES

For AES system the standard evaluation metric is the Quadratic Weighted Kappa (QWK) (Cohen, 1968), this measures the agreement between a models predictions and human scores. QWK is a quadratic weighted version of Cohen's Kappa, which accounts for the ordinal nature of essay scores. It ranges between -1 and 1, where 1 indicates perfect agreement, 0 indicates no agreement, and negative values indicate worse than random agreement. QWK is particularly useful in AES as it penalises larger discrepancies between predicted and actual scores more heavily than smaller ones.

In addition to QWK, regression metrics are very commonly used such as the Mean Squared Error (MSE) and Root Mean Squared Error (RMSE). These metrics provide an insight into the magnitude of predictive errors and a highly beneficial in continuing predictive tasks. Mean Absolute Error (MAE) is also occasionally used due to its resistance to outliers, though it lacks the sensitivity of QWK to ordinal score differences.

Some researchers report accuracy, especially in the form of accuracy (i.e., within ± 1 of the human score), to reflect more lenient scoring expectations similar to inter-rater variability among humans (Shermis and Burstein, 2013a; Taghipour and Ng, 2016). However, accuracy alone may not fully reflect a model's ranking reliability or nuanced error distribution, making it less preferred in rigorous evaluations (Williamson, Xi and Breyer, 2012; Yannakoudakis, Briscoe and Medlock, 2011).

Ultimately the combination of QWK, MSE and correlation coefficients is typically used to assess various aspects of a models performance. The multi-metric approach aims to ensure that AES models are evaluated not just on pure agreement but additionally on how well they can replicate the scoring patters of their human counterparts across a range of tasks.

2.6 Challenges in AES

Despite the significant progression in AES technology, several significant challenges remain. One major concern is bias and fairness within the scoring. AES models can learn and perpetuate biases present within training data, leading to an unfair scoring for certain groups of students. Recent research has scrutinized whether AES systems favour or penalize essays based on factors unrelated to the writings quality. Studies have found that a number of algorithms have exhibited different types of biases. For example, a comparative evaluation by (Litman, Baffour and Crossley, 2024). found that some AES models' scores varies systematically with students race and gender demographics, even when essay quality was consistent. These biases often occur when a model picks up on linguistic styles or topics more common for certain groups, or if the training essays are not properly balanced across demographics. This poses a risk where if these

issues are ignores, AES systems may reinforce educational inequalities by giving lower scores to those essays written by non-native speakers or students from under-represented backgrounds (Blodgett et al., 2020). Addressing these biases is challenging due to the requirement of identifying the subtle correlations and ensuring the models predictions are based purely on the relevant indicators. Ongoing research has including techniques such as bias audits, de-biasing models via data augmentations, fairness constraints and developing new "fairness" metrics to evaluate across subgroups of students. The assurance of AES fairness and lack of bias is an active concern within research as any real-world deployment of AES systems will requires high level of trust and accountability.

A second persistent challenge is cross-domain generalization (meaning a models ability to maintain its performance across new prompts and varying writing tasks outside of its training data). Traditionally, AES systems have been prompt-specific, an individual model developed trained for a singular essay prompt, trained on specific data. This approach is naturally limited in scalability as, requiring an individual model per task, and leads to issues in model robustness. In a practical ideal, AES systems would be capable of generalizing to a prompt of which it has never seen, without requiring further training. Research has shown that a model trained on one prompt is applied to an unseen prompt can lead to significant drops in performance, with some models only achieving 50% of their original QWK scores (Blanchard et al., 2013). The features and patterns learnt in training may not be applicable to a new prompt, due to the differences in vocabulary, structure and content required for different writing tasks. This problem of prompt dependency lead to a period of stagnation in AES research through the early 2000s, however with modern deep learning, a growing effort to tackle issues has resurfaced. Some approaches treat cross-prompt scoring as a domain adaption problem, where each prompt is a "domain" (Litman, Baffour and Crossley, 2024). Other methods inject prompt information into the model to aid in an AES systems adjusting its expectations for content (Blanchard et al., 2013). Some recent studies have reported progression with training prompt-agnostic models or multi-prompt training to improve generalization. The problem of a reliable cross-prompt AES remains unsolved - most models still perform best when fine-tuned on the target prompt data, to ensure a continent score criteria across problems is an open challenge.

Similarly there is the issue of a models ability to generalize across different proficiency levels and contexts of writing. Models are typically trained on a somewhat narrow band of essay types. Therefore when presented with an essay differing from their profile, their performance suffers. For instance, a system trained on high school essays may struggle when presented with elementary-level writing, or vice versa. Similarly as mentioned, scoring essays written by English language learners (ELLs/ESL students) is a challenge if the model has seen mostly native-level writing. Studies have noted that the widely-used ASAP corpus consists of essays by native English speakers in a timed exam context, which means models tuned to it may not handle non-native grammar errors or unusual expressions well. In contrast to this, real-world applications must deal with a full range of second-language writing issues. Another important aspect is the robustness to varied essay lengths or formats, in some cases essay length can impact grade (Li and Ng, 2024), and a model may unfairly rely on length if not carefully regularized. All these factors underline the need for AES systems of which are adaptable to different populations and conditions. Researchers are exploring techniques such as adversarial training, or building ensemble/hybrid models which combine neural scoring and rule-based checks to handle out of scope cases.

Nevertheless, ensuring an AES maintains its accuracy across any writer, across varying prompts, expertise level, genre and languages remains to be a long-term challenge. In multilingual settings this challenge poses an even greater struggle, a model trained in English cannot directly score in French or Spanish without significant alterations. As such cross-lingual AES is largely un-researched, due to the intricate language specific syntax difference. Sine process has been made in curating datasets for other languages however not to a significant enough level to allow for significant developments.

2.7 Gaps in Existing Research

While the recent advancements in AES have produces substantial performance improvements, several critical caps remain in modern research in areas limiting real world applicability.

One notable limitation is due to the lack of generalisability across prompt and writing genres. Many systems are prompt-specific, requiring separate methods for each essay tasks. This process restricts model scalability and makes it very challenging to deploy AES systems in a real educational setting where writing tasks vary substantially. While recent prompt-agnostic approaches have shown promising, they remain somewhat unexplored and are often only able to be evaluated on limited datasets.

Another constant challenge is the limited focus on fairness and bias (Litman, Baffour and Crossley, 2024). Existing models have shown to underperform of essays written by non-native speakers or students from background less represented in training datasets. Despite of the availability of datasets like TOEFL11 that include demographical metadata, limited studies make use of such data to evaluate and/or mitigate the models bias. This represents an important area for future work, particularly when given importance of bias free models in an educational setting.

Many current AES systems also only produce a singular score with little available insight into the specific strengths or weaknesses of the students work (Landauer, Laham and Foltz, 2003). There is a growing need for models to be capable of scoring off rubrics (as commonly used in university and school assessments), which would allow for more meaning full feedback to the student. Although some datasets support this type of scoring, these systems remain relatively underdeveloped compared to their holistic counterparts.

Additionally, current research focuses almost solely on types, English text, leaving a significant gap in assessments taken through handwritten responses or spoken essays. Expanding current systems to support handwriting or speech could make systems more inclusive and better aligned with real-world classroom use, especially in areas where there is lower access to technology.

Lastly, while many recent systems are achieving a high scoring accuracy, they often lack any form of interpretability. This raises concerns about transparency and trust in their decisions, as mentioned this is a significant consideration when used in a high-stakes educational context.

This project aims to address several of these gaps by developing a transformer-based AES system with the potential for future expansion into bias-free training, rubric-based scoring, and multi-modal input handling.

Chapter 3

Data Description and Preprocessing

3.1 Dataset Overview

The primary dataset being used for this project is the **Automated Student Assessment Prize (ASAP)** dataset, released by the Hewlett Foundation in 2012 as part of a public kaggle competition with the goal of advancing AES research. The dataset consists of almost 13000 essays written by students in response to eight different prompts, each with a unique scoring rubric. The prompts vary in genre, include argumentative, narrative and source based tasks thus providing a diverse set of writing styles and topics. The essays are scored on a scale of 1-6, with multiple human raters providing scores for each essay. This dataset has become a standard benchmark for AES systems, allowing for the comparison of different models and approaches.

The Essay lengths range from approximately 150 to 600 words, with each response being graded by 2 independent human raters. The scoring scales differ across different prompts, for example with some using a 1-6 scale and others using a 0-3 scale for example, depending on the complexity of the task.

Each record includes the following attributes:

- Unique essay ID
- Prompt ID
- Essay text
- Human scores
- Additional features such as the rater ID and scoring rubrics

The ASAP dataset has become a benchmark in the AES research community due to its relatively large size, accessibility, and variety of prompt types. It supports both prompt-specific modelling (where separate models are trained for each prompt) and prompt-agnostic modelling (where a single model is trained across all prompts). This flexibility makes it suitable for evaluating model generalisability — a key concern in the development of fair and scalable AES systems.

3.2 Dataset Attributes and Scoring Schemes

Each essay in the ASAP dataset is associated with one of the eight distinct prompts, each corresponding to a unique writing task. These prompts vary in genre, from narrative to persuasive and source-based writing, and differ in their complexity and structure. As a result, each prompt is scored according to its specific rubric designed to evaluate key aspects such as organization, grammar, coherence, creativity, or use of supporting evidence.

The available scores vary across the different prompts. Some prompts have limited possible marks (e.g., 0–3 or 0–6), while others have broader possibility such as 0–12 or 0–60. These scales are not directly comparable, as the same numeric score may imply different levels of proficiency depending on the prompt. This discrepancy requires label normalization strategies to enable rapid cross-prompt training and fair evaluation, which are discussed in Section.

Most essays in the dataset have been scored by two trained human raters. Where both scores are available, With the average typically being used for the "final grade". There are a number of essays where only one mark is present, while this may lead to a lower quality of training due to an inherent risk of higher marking bias, they will be included.

Each essay record also contains additional metadata such as the prompt ID, the identifier of the essay set, and the number of rating agents. While these attributes are not used as direct model inputs, they are essential for organising the data, constructing prompt-specific training splits, and interpreting performance across different writing tasks.

3.3 Data Selection and Filtering

In order to ensure that our data is to a high quality and consistent. Filters were applied to the ASAP in order to remove cases of missing fields, issues with formatting and/or corrupted characters that would interfere with the tokenisation processes.

Similarly any essays shorter than 50 words were filtered out, this is due to them being too shorted to contain sufficient context for a meaningful score and posing a risk to increasing noise during training. As mentioned those with only one rater were still included however any cases where no score is present, the item was omitted.

No additional sampling or balancing techniques were used at this stage. The distribution of prompts were kept as original in order to maintain a consistent representation of each tasks. The resulting dataset is structurally consistent and ready for downstream preprocessing and model training.

3.4 Tokenisation and Text Preprocessing

Tokenisation and text preprocessing will follow standard practices for transformer-based language models, with particular alignment to the requirements of the selected model architecture. The hugging face version of WordPiece tokenizer will be used to segment essay text into subword units. This approach is taken due to its robustness to spelling errors and uncommon works, of which is frequently found in students writing

Minimal additional preprocessing is expected as the models are typically trained on raw preprocessed text. Some small changes may be made should large improvements be found in

the models performance. However, the basic cleaning procedures will be carried out to ensure compatibility with the tokenisation pipeline.

Essay length has also be considered, focusing on model input constraints. Essays exceeding the model's maximum input length (typically 512 tokens) will be truncated, and those far below expected length thresholds will be flagged during data inspection. Any preprocessing strategies will be consistently applied across all training, validation, and test splits.

3.5 Label Normalisation

In order to deal with the variation in scoring scales across essays, label normalisation will be applied to ensure a consistent scale during training. This is important in prompt-agnostic models where a singular model will struggle to generalise in cases with varying scores.

The decided method is min-max normalisation, in which each essays original score is scaled to fall within the boundaries of [0,1], based off the minimum and maximum available for each prompt. This allows the the model to understand the relative performance of an essay, irrespective of the prompt-specific scoring ranges.

Within the evaluation processes model outputs will be rescaled back to the original score ranges to aid with interpretability and for comparisons with baseline methods. This normalisation process ensures that the training will remain consistent while each score remains as a meaningful representation of students performance.

Alternatives such as z-score and ordinal regression may be explored should the initial results suggesting our current method is ineffective. However, min-max scaling remains most likely due to the straightforward and interpretable process.

3.6 Limitations and Considerations

As mentioned the ASAP dataset acts as a widely used benchmark for AES research, it has a number of limitations should be acknowledged/addressed. These factors may influence a models generalisability and will inform the design decision in later stages of the project.

A key issue is that the dataset only contains typed essays written in English taken from specific educational contexts. As a result, it may not generalise to handwritten, multilingual or submissions where students are from different academic and cultural backgrounds. This likely restricts the models applicability and limits the diversity of writing styles and linguistic structures represented in training.

Additionally, each prompt has an independent scoring rubric and scale. Although the label normalisation processes will align the ranges, the respective scoring criteria may vary significantly across the prompts. This introduces a significant degree of inconsistency that may affect model training, particularity in prompt-agnostic setting.

The distribution of essays across each prompt is also uneven. Some score levels are under-represented, increasing the risk of a model bias towards more frequent outcomes. This imbalance can be relevant when working with regression-based models or when interpreting model predictions made near the extremes of the scoring scales.

Lastly, while the majority of essays have been scored by two rates, the dataset does not include any details with their identity or agreements between scores beyond the raw score. This limits the ability to assess rater reliability and thus correct for any potential systematic biases within the labelling

These limitations do not prevent the development of effective models, but they do highlight the need for careful evaluation, clear methodology and the potential inclusion of additional datasets where necessary.

3.7 Additional datasets for model extensions

In addition to the ASAP dataset, a number of additional datasets have been considered to support the planned extensions onto the base ASE model. These additional resources are intended to aid in the implementation of bias analysis, multi-modal scoring, multi-linguistic scoring and feedback generation (among other potential additions).

One key option is the **TOEFL11 corpus**, this dataset is comprised of a series of non-native English speaker essays from eleven different language backgrounds. This dataset contains metadata of the writers native language, gender and English proficiency level. This addition enables exploration into demographic bias and the models ability to generalise across the writing patterns of different groups(Blanchard et al., 2013).

The **First Certificate in English (FCE) Dataset** produced from the Cambridge learner Corpus, offers a different approach. Unlike the ASAP, the FCE gives detailed annotations assessing the essays grammar, coherence, vocabulary, and structure. This aid with multi-dimensional scoring and enables additions focusing on explainability and providing targetted feedback.

To support extensions into handwritten essay scoring, the **IAM Handwriting Database** is being considered (Marti and Bunke, 2002). This dataset consists of handwritten English text collected from over 600 writers. While this dataset has little to do with the domain of AES models it will provide value to the models use cases through enabling the comprehension of hand written text. As previously mentioned, much of students work is completed by hand and therefore including an implementation for processing handwritten text is a key requirement for real-world implementation.

Chapter 4

Methodology

This section outlines the methodological framework used to develop, train and evaluate the AES system. This methodology serves as both a practical guide for implementation and as a transparent report to aid with reproducibility. The approach blends modern NLP techniques with robust evaluation strategies in order to address key limitations within the current research - specifically related to generalisability, interpretability, and bias.

There are six key stages: data preprocessing, baseline model implementation, transformer-based model fine-tuning, evaluation metric selection, explainability analysis and fairness evaluation. Each stage is designed to address the project's research questions, including comparative performance analysis, interpretability and demographic fairness.

All code and experimental steps were developed in keeping with the goal of transparency and reproducibility in mind, utilising open source libraries, publicly available resources (where possible) and making all code available on the Git repository. Ethical considerations, including bias analysis and fairness audits are integrated through the experimental design.

4.1 Research Design and Approach

This project adopts an experimental research design grounded in applied natural language processing. The primary objective is to develop and evaluate a transformer-based Automated Essay Scoring (AES) system, comparing its performance, fairness, and interpretability against traditional machine learning baselines. The approach is comparative and quantitative in nature, with a strong emphasis on reproducibility and ethical evaluation.

The study is structured into six methodological stages: (i) data preprocessing, (ii) baseline model implementation, (iii) transformer model fine-tuning, (iv) metric-based performance evaluation, (v) model explainability analysis, and (vi) fairness auditing. Each stage is aligned with the project's research questions to ensure that the design supports both technical innovation and responsible AI principles.

Transformer-based models (specifically RoBERTa) were chosen due to their strong contextual representation capabilities and demonstrated performance in AES tasks. Traditional models such as ridge regression and support vector machines are implemented as benchmarks to contextualise improvements in predictive accuracy. Additionally, the study integrates explainability tools

(e.g. SHAP) and fairness metrics (e.g. demographic parity and QWK gaps) to assess model transparency and ethical robustness.

Overall, the research design balances model performance with interpretability and inclusivity, providing a systematic framework for both evaluation and potential real-world deployment in educational contexts.

4.2 Data Collection and Preprocessing

4.2.1 Dataset Selection and Description

The dataset selected for this project is the Automated Student Assessment Prize (ASAP) corpus, originally released by the Hewlett Foundation as part of a public Kaggle competition in 2012. It contains approximately 13,000 student essays written in response to eight distinct prompts, each accompanied by its own scoring rubric and assessed by one or two trained human raters. The dataset spans a variety of genres, including narrative, argumentative, and source-based essays, which provides a strong foundation for evaluating the generalisability of automated scoring models.

Each essay entry includes a unique ID, the prompt identifier, the full essay text, and the assigned human scores. Scoring scales vary by prompt, ranging from narrow intervals (e.g. 0-3 or 0-6) to broader ones (e.g. 0-12 or 0-60), depending on the rubric complexity. This variability supports both prompt-specific and prompt-agnostic modelling but also necessitates appropriate score normalisation strategies.

The ASAP dataset was chosen due to its widespread use as a benchmark in AES research, the diversity of writing tasks it encompasses, and its suitability for evaluating both performance and fairness in transformer-based scoring models.

4.2.2 Tokenisation and Cleaning

To prepare the dataset for model training, a standardised text preprocessing pipeline was applied. Essays with missing text fields, invalid characters, or structural corruption were removed. Additionally, responses under 50 words were excluded to eliminate noise and ensure sufficient content for meaningful scoring.

Tokenisation was carried out using the HuggingFace implementation of the WordPiece tokenizer, specifically configured for compatibility with RoBERTa. This approach segments text into subword units, offering robustness to typographical errors, misspellings, and rare words—features often present in student writing.

No stop word removal, stemming, or lemmatisation was performed, as these steps are unnecessary for transformer models. Essays exceeding the model's maximum input length of 512 tokens were truncated from the end, while shorter texts were padded at the batch level. These decisions maintain compatibility with transformer constraints while preserving important semantic content, particularly at the beginning of the text.

4.2.3 Label Normalisation

Because each prompt in the ASAP dataset uses a different scoring scale, a label normalisation strategy was required to enable uniform training across prompts. This project applies min-max normalisation, scaling each essay's score to a standardised range of [0, 1] based on the minimum and maximum values for that prompt.

This method allows a single model to generalise over multiple prompts without overfitting to prompt-specific scoring intervals. During evaluation, predicted scores are rescaled to their original range to facilitate interpretability and comparison against human-marked scores and baseline models.

Alternative techniques such as z-score normalisation and ordinal regression were considered but deemed less appropriate given the non-Gaussian distribution and ordinal nature of essay scores. Min-max scaling was ultimately selected due to its simplicity, interpretability, and effectiveness in existing AES literature.

4.3 Model Architecture and Implementation

4.3.1 Baseline Models

To establish a comparative baseline for the performance of the transformer-based AES system, several traditional machine learning regressors were implemented. These include, Ridge Regression, Support Vector Regressors (SVR), and a Random Forrest Regressor. All baseline models operate on the TF-IDF representations of essays, which were selected due to their ability to preserve frequency-weighted term information, while remaining interpretable and computational efficient (Zhang and Litman, 2020).

The TF-IDF vectorizer was configures with a maximum of 10,000 features and allowed for unigrams and bigrams (`i.e., ngram_range(1,2)`), enabling the capture of both single word frequencies and short collocations. Following this, essay texts are split into training, validation and testing split(80%,10%,10% respectively) using a random seen to ensure reproducibility. The target variable was the win-max normalized essay score, as discussed in section 4.2.3.

The ridge regression model was selected due to its regularized nature, which controls overfitting through the penalization of large weights. This served as the primary linear baseline. SVR with a radial basis function kernel was used to introduce non-linear modeling, allowing for the capture of more complex relationships between essay text and score. The random forrest regressor, as ensemble of decision trees, provides a thrid approach by aggregating multiple tree-based leraners, of which are robust to outliers and can modle the non linear feature interactions without requireing scaling (Breiman, 2001).

Each model was evaluated with MSE and QWK - the latter being especially relevant for AES due to its sensitivity to the ordinal label discrepancies (Cohen, 1968). To compute QWK, predicted and actual scores were rescaled to match the typically 0-12 range of AES evaluations (Taghipour and Ng, 2016).

The baseline models serve to establish reference points for comparing the effectiveness of transformer-based methods in both predictive accuracy and fairness, as will be discussed in chapter 5.

Chapter 5

Experiments and Results

5.1 Baseline Performance

5.2 Transformer Performance

5.3 Error Analysis

5.4 Model Explainability

5.5 Fairness and Bias Analysis

5.6 Comparison with Human Raters

Chapter 6

Discussion

6.1 Interpretation of Results

6.2 Contributions to the Field

6.3 Limitations

6.4 Implications for Educational Technology

Chapter 7

Conclusion

Appendix A

Design Diagrams

Appendix B

User Documentation

Appendix C

Raw Results Output

Appendix D

Code

D.1 File: 01 Data Imports and Preprocessing

```
#!/usr/bin/env python
# coding: utf-8

# #  Data Preprocessing: ASAP Dataset
#
# This notebook holds all of the preprocessing steps required for
    the projects relevant datasets. It includes:
# - Loading and inspecting the original ASAP data
# - Cleaning and filtering essay entries
# - Handling scoring ranges and label normalisation
# - Preparing tokenised inputs for model training
#
# All steps in this notebook follow the methodology described in
    Chapter 4 of the dissertation and are designed to ensure
    reproducibility and compatibility with the HuggingFace
    transformer framework.
#

# ##  Loading the ASAP Dataset
#
# The following section loads the original Automated Student
    Assessment Prize (ASAP) dataset from the "raw" data directory.
    The dataset includes approximately 13,000 student essays
    across eight distinct writing prompts, each with unique
    scoring rubrics.
#
# Key columns of interest include:
# - `essay_id`: unique identifier for each essay
# - `essay_set`: the prompt ID (18)
# - `essay`: the full essay text
# - `domain1_score`: the primary human-assigned score
    ######should this also include the rest of the scoring, why am
    I just blindly trusting the domain1_score
#
# Additional fields (e.g., `rater_1_domain1`, `rater_2_domain1`)
    may be used for advanced analysis but are not essential to the
    initial modelling phase.
#

# In[1]:


import pandas as pd

# Load in the ASAP dataset - Rel3 is used as it is the most up to
    date and contains the least errors
df = pd.read_csv("../data/raw/asap-aes/training_set_rel3.tsv",
    sep='\t', encoding='latin1')

# Basic info - for checks
print(f"␣Loaded␣dataset␣with␣shape:␣{df.shape}")
print("\n␣Column␣names:")
print(df.columns.tolist())

# Preview a few entries
df[['essay_id', 'essay_set', 'domain1_score', 'essay']].head()


# In[7]:


# Drop any rows with missing essays or scores
df = df.dropna(subset=['essay', 'domain1_score'])

# Strip whitespace and remove essays with too few words
df['essay'] = df['essay'].str.strip()
df['word_count'] = df['essay'].apply(lambda x: len(x.split()))
df = df[df['word_count'] >= 50]  # remove very short essays

# Reset index after filtering
df = df.reset_index(drop=True)

# Show updated shape and word count stats
print(f"␣Cleaned␣dataset␣shape:␣{df.shape}")
print(df['word_count'].describe())


# In[8]:


# Get per-prompt score ranges
prompt_stats =
    df.groupby('essay_set')['domain1_score'].agg(['min',
    'max']).rename(columns={'min': 'min_score', 'max':
    'max_score'})
print("␣Score␣ranges␣by␣prompt:")
print(prompt_stats)

# Merge these stats back into the main DataFrame
```

```
df = df.merge(prompt_stats, left_on='essay_set', right_index=True)

# Apply MinMax normalisation
df['score_scaled'] = (df['domain1_score'] - df['min_score']) /
    (df['max_score'] - df['min_score'])

# Preview scaled scores
df[['essay_id', 'essay_set', 'domain1_score',
    'score_scaled']].head()


# In[9]:
```

```
#Saving and Visualize the cleaned dataset
df.to_csv("../data/Processed/asap_cleaned.csv", index=False)
import matplotlib.pyplot as plt
df['score_scaled'].hist(bins=30)
plt.title("Distribution␣of␣Normalised␣Scores")
plt.xlabel("score_scaled")
plt.ylabel("Count")
plt.show()


## ^^ that looks like a lot of full mark scores? am I happy with
    that? ^^

#
```

D.2 File: 02 Tokenization

```
#!/usr/bin/env python
# coding: utf-8

# #  Tokenisation for Transformer-Based AES
#
# This notebook prepares the cleaned ASAP dataset for input into
    transformer-based models by applying the RoBERTa tokenizer. It
    includes:
#
# - Loading the normalised, cleaned dataset
# - Initialising the RoBERTa tokenizer
# - Applying padding and truncation
# - Outputting a tokenised HuggingFace-compatible dataset
#
# Tokenisation aligns with the preprocessing strategy described in
    Section 4.2 of the dissertation.
#

# In[4]:


import pandas as pd
from transformers import RobertaTokenizerFast
from datasets import Dataset


# In[ ]:


# Load your cleaned and normalised dataset
df = pd.read_csv("../data/processed/asap_cleaned.csv")

# Check a sample
df[['essay_id', 'essay_set', 'essay', 'score_scaled']].head()


# In[ ]:


# Load the RoBERTa tokenizer
tokenizer = RobertaTokenizerFast.from_pretrained("roberta-base")

# Set max length for tokenisation
MAX_LENGTH = 512


# In[ ]:


# HuggingFace Datasets expects columns as dictionary entries
dataset = Dataset.from_pandas(df[['essay', 'score_scaled']])  #
    only keep needed fields

# Tokenisation function
def tokenize_function(example):
    return tokenizer(
        example['essay'],
        padding="max_length",
        truncation=True,
        max_length=MAX_LENGTH,
    )

# Apply tokenizer across dataset
tokenised_dataset = dataset.map(tokenize_function, batched=True)


# In[ ]:


from datasets import DatasetDict

# Split into 80% train, 10% val, 10% test
split_dataset = tokenised_dataset.train_test_split(test_size=0.2,
    seed=42)
val_test_split =
    split_dataset['test'].train_test_split(test_size=0.5, seed=42)

# Combine into DatasetDict
dataset_dict = DatasetDict({
    'train': split_dataset['train'],
    'validation': val_test_split['train'],
    'test': val_test_split['test']
})

# Check sizes
print(dataset_dict)


# In[ ]:


dataset_dict.save_to_disk("../data/processed/tokenised_asap_split")
```

D.3 File: 03 Baseline Models

```
#!/usr/bin/env python
# coding: utf-8

# In[1]:


import pandas as pd
import numpy as np
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.linear_model import Ridge
from sklearn.svm import SVR
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error
from sklearn.preprocessing import MinMaxScaler

from datasets import load_dataset
from tqdm.notebook import tqdm

# Optional: for QWK
get_ipython().system('pip install --quiet scikit-learn scipy')
from sklearn.metrics import cohen_kappa_score

# Load your cleaned CSV
df = pd.read_csv("../Data/Processed/asap_cleaned.csv")
df.head()


# In[3]:


# Use 'essay' column as input text
texts = df['essay'].astype(str).values

# Use normalised score as target (assumes already scaled to [0, 1])
labels = df['score_scaled'].values


# In[4]:


# First split: Train + Temp (for val/test)
X_train, X_temp, y_train, y_temp = train_test_split(
    texts, labels, test_size=0.2, random_state=42)

# Second split: Temp  Validation + Test
X_val, X_test, y_val, y_test = train_test_split(
    X_temp, y_temp, test_size=0.5, random_state=42)

print(f"Train size: {len(X_train)}")
print(f"Validation size: {len(X_val)}")
print(f"Test size: {len(X_test)}")


# In[5]:


vectorizer = TfidfVectorizer(max_features=10000, ngram_range=(1,
    2))

X_train_vec = vectorizer.fit_transform(X_train)
X_val_vec = vectorizer.transform(X_val)
X_test_vec = vectorizer.transform(X_test)


# In[6]:


ridge = Ridge(alpha=1.0)  # You can tune alpha later if desired
ridge.fit(X_train_vec, y_train)

# Predict on validation and test
val_preds_ridge = ridge.predict(X_val_vec)
test_preds_ridge = ridge.predict(X_test_vec)


# In[7]:


def qwk(y_true, y_pred, min_rating=0, max_rating=1):
    """
    Quadratic Weighted Kappa. Assumes inputs scaled to [0, 1].
    For scoring purposes, predictions are mapped back to 0-12
    scale (ASAP-style).
    """
    y_pred_rounded = np.round(y_pred * 12).astype(int)
    y_true_rounded = np.round(y_true * 12).astype(int)
    return cohen_kappa_score(y_true_rounded, y_pred_rounded,
        weights="quadratic")

mse_ridge = mean_squared_error(y_test, test_preds_ridge)
qwk_ridge = qwk(y_test, test_preds_ridge)
```

```
print(f"Ridge␣Regression␣-␣MSE:␣{mse_ridge:.4f},␣QWK:␣
    {qwk_ridge:.4f}")


# In[8]:


svr = SVR(kernel="rbf", C=1.0, epsilon=0.1)
svr.fit(X_train_vec, y_train)

# Predict
val_preds_svr = svr.predict(X_val_vec)
test_preds_svr = svr.predict(X_test_vec)


# In[9]:


mse_svr = mean_squared_error(y_test, test_preds_svr)
qwk_svr = qwk(y_test, test_preds_svr)

print(f"Support␣Vector␣Regression␣-␣MSE:␣{mse_svr:.4f},␣QWK:␣
    {qwk_svr:.4f}")


# In[10]:


rf = RandomForestRegressor(
    n_estimators=100,        # Number of trees
    max_depth=None,          # You can limit this to avoid
        overfitting
    random_state=42,
    n_jobs=-1                # Use all available cores
)
rf.fit(X_train_vec, y_train)

# Predict
val_preds_rf = rf.predict(X_val_vec)
test_preds_rf = rf.predict(X_test_vec)


# In[11]:


mse_rf = mean_squared_error(y_test, test_preds_rf)
qwk_rf = qwk(y_test, test_preds_rf)

print(f"Random␣Forest␣Regressor␣-␣MSE:␣{mse_rf:.4f},␣QWK:␣
    {qwk_rf:.4f}")


# In[12]:


results_df = pd.DataFrame({
    "Model": ["Ridge␣Regression", "Support␣Vector␣Regression",
        "Random␣Forest␣Regressor"],
    "MSE": [mse_ridge, mse_svr, mse_rf],
    "QWK": [qwk_ridge, qwk_svr, qwk_rf]
})

# Round for cleaner display
results_df = results_df.round(4)
display(results_df)
```

D.4 File: 03 Baseline Models

```
#!/usr/bin/env python
# coding: utf-8

# In[ ]:


from datasets import load_from_disk

# Load tokenised dataset
dataset_dict =
    load_from_disk("../data/processed/tokenised_asap_split")

# Make sure labels column is correct
if "score_scaled" in dataset_dict["train"].features:
    dataset_dict = dataset_dict.rename_column("score_scaled",
        "labels")


# In[2]:


from transformers import RobertaForSequenceClassification,
    TrainingArguments

# Load RoBERTa with regression head
model =
    RobertaForSequenceClassification.from_pretrained("roberta-base",
    num_labels=1)

training_args = TrainingArguments(
    output_dir="./results",
    evaluation_strategy="epoch",
    save_strategy="epoch",
    learning_rate=2e-5,
    per_device_train_batch_size=8,
    per_device_eval_batch_size=8,
    num_train_epochs=3,
    weight_decay=0.01,
    logging_dir="./logs",
    logging_steps=10,
    load_best_model_at_end=True,
    metric_for_best_model="eval_loss",
    save_total_limit=2
)
```

```
# In[3]:


import evaluate
import numpy as np

mse_metric = evaluate.load("mse")
r2_metric = evaluate.load("r_squared")

def compute_metrics(eval_pred):
    predictions, labels = eval_pred
    predictions = predictions.squeeze()
    labels = labels.squeeze()
    mse = mse_metric.compute(predictions=predictions,
        references=labels)
    r2 = r2_metric.compute(predictions=predictions,
        references=labels)
    return {
        "mse": mse["mse"],
        "r2": r2["r_squared"]
    }


# In[5]:


from transformers import Trainer

trainer = Trainer(
    model=model,
    args=training_args,
    train_dataset=dataset_dict["train"],
    eval_dataset=dataset_dict["test"],
    tokenizer=None,
    compute_metrics=compute_metrics,
)

trainer.train()
eval_results = trainer.evaluate()
print("Evaluation Results:", eval_results)


# In[ ]:
```

Bibliography

Attali, Y. and Burstein, J., 2006. Automated essay scoring with e-rater® v.2. *The journal of technology, learning and assessment* [Online], 4(3). Available from: https://ejournals.bc.edu/index.php/jtla/article/view/1647.

Beigman Klebanov, B., Burstein, J., Gyawali, B. and Sabatini, J., 2021a. Limitations of automated essay scoring when evaluating narrative writing. *Assessing writing* [Online], 47, p.100511. Available from: https://doi.org/10.1016/j.asw.2020.100511.

Beigman Klebanov, B., Madnani, N., Cahill, A. and Flor, M., 2021b. Limitations of automated essay scoring when evaluating narrative writing. *Assessing writing* [Online], 47, p.100511. Available from: https://doi.org/10.1016/j.asw.2020.100511.

Blanchard, D., Tetreault, J., Higgins, D., Cahill, A. and Chodorow, M., 2013. Toefl11: A corpus of non-native english. *Proceedings of the eighth workshop on innovative use of nlp for building educational applications*. pp.45–52.

Blodgett, S.L., Barocas, S., Daumé III, H. and Wallach, H., 2020. Language (technology) is power: A critical survey of "bias" in nlp. *Proceedings of the 58th annual meeting of the association for computational linguistics* [Online], pp.5454–5476. Available from: https://aclanthology.org/2020.acl-main.485/.

Breiman, L., 2001. Random forests. *Machine learning*, 45(1), pp.5–32.

Burstein, J., Chodorow, M. and Leacock, C., 2003. Automated essay evaluation: The criterion online writing evaluation service. *Ai magazine*.

Cohen, J., 1968. Weighted Kappa: Nominal scale agreement provision for scaled disagreement. *Psychological bulletin*, 70.

Devlin, J., Chang, M.W., Lee, K. and Toutanova, K., 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *arxiv preprint arxiv:1810.04805* [Online]. Available from: https://arxiv.org/abs/1810.04805.

Dikli, S., 2006. An overview of automated scoring of essays. *The journal of technology, learning and assessment* [Online], 5(1). Available from: https://ejournals.bc.edu/index.php/jtla/article/view/1646.

Kaggle, 2012. Automated student assessment prize (asap) aes dataset [Online]. Available from: https://www.kaggle.com/competitions/asap-aes/data.

Landauer, T.K., Laham, D. and Foltz, P.W., 2003. Automated scoring and annotation of essays with the intelligent essay assessor [Online]. *Automated essay scoring: A cross-disciplinary perspective*. Mahwah, NJ: Lawrence Erlbaum Associates, pp.87–112. Available

from: https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.83.3597&rep=rep1&type=pdf.

Li, S. and Ng, V., 2024. Automated essay scoring: A reflection on the state of the art [Online]. *Proceedings of the 2024 conference on empirical methods in natural language processing (emnlp)*. Available from: https://aclanthology.org/2024.emnlp-main.991.

Litman, D., Baffour, P. and Crossley, S., 2024. Fairness in automated essay scoring: A comparative analysis of shallow and deep learning algorithms [Online]. *Proceedings of the 19th workshop on innovative use of nlp for building educational applications (bea)*. Association for Computational Linguistics, pp.184–194. Available from: https://aclanthology.org/2024.bea-1.18.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V., 2019. Roberta: A robustly optimized bert pretraining approach. *arxiv preprint arxiv:1907.11692* [Online]. Available from: https://arxiv.org/abs/1907.11692.

Lundberg, S.M. and Lee, S.I., 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems* [Online], 30, pp.4765–4774. Available from: https://arxiv.org/abs/1705.07874.

Marti, U.V. and Bunke, H., 2002. The IAM-database: An english sentence database for offline handwriting recognition. *International journal on document analysis and recognition*, 5.

Page, E.B., 1966. The imminence of grading essays by computer. *Phi delta kappan* [Online], 47(5), pp.238–243. Available from: https://wac.colostate.edu/docs/books/usu/machine/chapter17.pdf.

Page, E.B. and Petersen, K.N., 2003. *The computer moves into essay grading: Updating the ancient test*. New York: Pearson.

Shermis, M. and Burstein, J., 2013a. *Handbook of automated essay evaluation: Current applications and new directions*. Routledge.

Shermis, M.D. and Burstein, J., 2013b. *Handbook of automated essay evaluation: Current applications and new directions* [Online]. Routledge. Available from: https://www.routledge.com/Handbook-of-Automated-Essay-Evaluation-Current-Applications-and-New-Directions/Shermis-Burstein/p/book/9780415810968.

Taghipour, K. and Ng, H.T., 2016. A neural approach to automated essay scoring [Online]. *Proceedings of the 2016 conference on empirical methods in natural language processing*. pp.1882–1891. Available from: https://aclanthology.org/D16-1193/.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L. and Polosukhin, I., 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Williamson, D.M., Xi, X. and Breyer, F.J., 2012. A framework for evaluation and use of automated scoring. *Educational measurement: Issues and practice*, 31(1), pp.2–13.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A. and Hovy, E., 2016. Hierarchical attention networks for document classification [Online]. *Proceedings of the 2016 conference of the*

north american chapter of the association for computational linguistics. Association for Computational Linguistics. Available from: `https://aclanthology.org/N16-1174`.

Yannakoudakis, H., Briscoe, T. and Medlock, B., 2011. A new dataset and method for automatically grading esol texts. *Acl*, pp.180–189.

Zhang, T. and Litman, D., 2020. Automated essay scoring: A survey of the state of the art. *Proceedings of the 28th international conference on computational linguistics*. pp.3739–3749.