CM50175 – Project Proposal

Oscar Dos Santos Nunes
Student ID: JOFDSN20

28 March 2025

Project Title

The current title for the project is **"Automated Essay Scoring Using Deep Learning with XXX"**.

This title reflects to general aim for the work to design, implement and evaluate and automated essay scoring (AES) algorithm focusing on the utilization of deep learning techniques.

However, the title will be updated as the project progresses and the research focus becomes more defined. The current placeholder "XXX" represents a yet to be decided add-on. This may involve the implementation of a specific deep-learning architecture—such as LSTM, BERT or a hybrid transformers model or the integration of additional functionality such as expandability, feedback generation, or the processing of handwritten text.

Some possible titles include:

- *Automated Assessment of Student Writing Using Deep Learning*
- *Exploring Transformer Architectures for Automated Essay Scoring*
- *Transformer-Based Models for Automated Essay Scoring*
- *From Grade to Guidance: Enabling Feedback Generation in AES Systems*
- *Beyond Scoring: Generating Feedback with Transformer-Based AES*
- *Expanding Automated Essay Scoring to Handwritten Responses Using Deep Learning*
- *Explainable AI for Essay Assessment: A Deep Learning Perspective*
- *Explainable Automated Essay Scoring Using Deep Neural Networks*

Problem Statement

Essay based work is an essential part of the educational process and therefore is vital for performance evaluation. However the grading of essays is a highly subjective process and can be time consuming. Studies have shown that the grading of essays by different teachers can vary significantly, with some teachers being more lenient than others. This can lead to inconsistencies in grading and can affect the overall performance of students. These issues are especially prevalent in larger-scale assessments (e.g. GCSEs, ALevels and Uni assessments) where many markers are involved. The use of automated essay scoring (AES) systems can help to address these issues by providing a more consistent and objective grading process.

AES systems aim to address these limitations through computational techniques to evaluate the quality of written text. Early AES systems such as Project Essay Grade, relied very heavily on manually engineered features such as grammar usage, word length, and sentence structure,

combined with traditional machine learning models. While these systems showed success (e-rater being used in english as a foreign language scoring), they often lacked robustness and generalization capabilities.

Recent advancements in deep learning and natural language processing have introduced more sophisticated models and approaches to AES. Neural architectures such as LSTM networks and transformers based models such as BERT have achieved high performance across a variety of NLP tasks, including text classification and sentiment analysis. For example, BERT achieved a GLUE benchmark score of 80.5 in its original implementation, showing its ability to to model contextual information and the relationships between words in a sentence. The models are well suited to AES tasks due to their ability to contextualize meaning beyond the surface-level features.

However, there are still several challenges to overcome. Deep networks often act as "black boxes", which offer very little interpretability into how they arrive at an output. This creases a concern for use in education where expandability/feedback and essential. Furthermore, these models require vast quantities of labeled-data of which is not always available. Additionally there are a number of documented risks regarding the algorithmic bias in AES particularly when dealing with non-native speakers or student with a less represented background than in the training data.

Moreover, much of the existing AES models focus solely on predicting a final holistic score, with less of a focus on providing additives that would be given by a human marker, such as feedback on specific aspects, annotated marking , or the ability to discuss the current mark and how to improve. Each of these would substantially improve the possible use of AES in education.

This project aims to explore and evaluate the application of deep learning techniques to automate essay scoring, which a focus on performance, fairness, and extensibility. Through the implementation of a deep learning model, the research aims to contribute to the development of a more accurate, explainable and general use AES system.

Objectives and Research Questions

This project aims to investigate the use of deep learning for automated essay scoring, focusing on the development of a model that is accurate, fair and explainable. The ultimate goal is the develop a system that can generate consistent and reliable scoring for comparable to that of human marking, while also considering the systems interpretability and the potential for future enhancements such as feedback generation or support for handwritten text

0.1 Objectives

1. **Review Existing AES Systems And literature:** *Identify the strengths, limitations, and gaps in current approaches, particularly in those using deep learning techniques.*

2. **Select And Prepare Datasets:** *Selecting, preparing and pre-processing datasets for training and evaluation of the model.*

3. **Develop And Compare Model Performance:** *Implementing and evaluating a deep learning model for AES, comparing its performance to traditional machine learning approaches.*

4. **Evaluate The Performance, Fairness And Interpretability:** *Investigate any issues with Biases, inconsistencies in the methods, as well as the models transparency and "trustworthiness".*

5. **Explore Potential Enhancements:** *Investigate the potential for future enhancements to the model, such as feedback generation, or multi-modal input.*

0.2 Research Questions

1. To what extent can transformer-based models outperform traditional machine learning approaches within the context of an automated essay scoring system?
2. What types of biases are present in AES systems, and how can deep learning models be adapted to mitigate these issues?
3. what are the potential trade offs between scoring accuracy, computational efficiency, and interpretability within the current AES techniques?
4. Can AES models by extended to provide meaningful and human-like feedback to support their usage within the educational setting?

Background and Related Work

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) and e-rater 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 students work.

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 the scalability and generalization 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 and RoBERTa have demonstrated a impressive results across ranging 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.

Despite these advancements, a number of obstacles still exists, deep learning models frequently act as "black boxes", providing little to no room for interpretation. This is a major drawback for AES, because user trust and educational integration rely on the reasoning behind a score. Second, in more recent years there is a growing concern about algorithmic bias, especially when it comes to under-represented student groups or non-native english speakers. These issues raise questions about fairness, generalisability, and the ethical use of AES in high-stakes contexts. 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 or attention-based visualizations have had lots of research recently to focus on improving the models transparency, while hybrid systems that combine neural and symbolic components have showed promise in improving 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 avenues for expansion.

Through comparing results against against traditional baselines and considering explainability and ethical implications, this work aims to contribute to all research seeking to make AES systems more robust, transparent, and educationally useful.

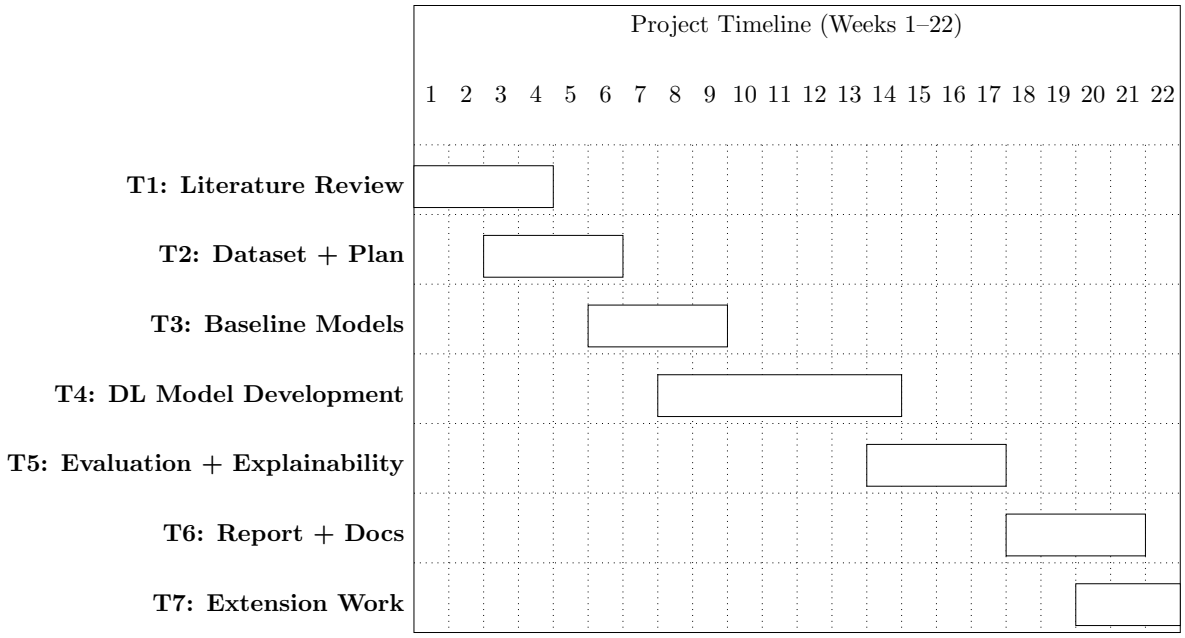
Project Plan

Task Breakdown

Bellow outlines the key tasks associated with the project, along with their expected duration for completion. Important to note that week 1 is considered as the week commencing March 31st.

Task ID	Task Description	Duration (Weeks)	Deliverables
T1	Literature review and intro sections	Weeks 1 - 4	Report sections completed and ready for submission
T2	Dataset pre-processing and model development plan	Weeks 3 - 6	Detailed plan for model development and datasets ready for next steps
T3	Baseline model performance evaluation	Weeks 6 - 9	Use related models and classic ML techniques to establish a baseline for performance
T4	Deep Learning model development	Weeks 8 - 14	Using related work and other models, develop out own AES system
T5	Evaluation; performance, fairness, and explainability	Weeks 14 - 17	Document models performance against key metrics and other models
T6	Reporting and documentation	Weeks 18 - 21	Complete the report and discuss the optional extensions
T7	System extension development	Weeks 20 - 22	Develop additional systems and discuss changes to performance

Timeline


Project Timeline (Weeks 1–22)
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
T1: Literature Review
T2: Dataset + Plan
T3: Baseline Models
T4: DL Model Development
T5: Evaluation + Explainability
T6: Report + Docs
T7: Extension Work

Dependencies and Risks

This projects tasks are sequential in nature, with each task building upon the findings of the previous one. due to that I have decided to omit a discussion for the dependencies. However below is an outline of the potential risks and the steps that will be taken to mitigate their effects.

Risk	**Impact**	**Mitigation Strategy**
Limited access to large, well-labelled datasets	May reduce model performance and generalisability	Use multiple publicly available datasets; explore data augmentation techniques; optionally explore synthetic data generation using GenAI
Inconsistent scoring methods and scales across datasets	May skew model results and reduce comparability across sources	Implement stratified sampling, resampling, or weighted loss functions to normalise scoring differences
Lack of experience with deep learning and transformer architectures	Poor model performance; longer development time	Allocate early project time to tutorials and documentation; replicate open-source AES models to build understanding
Deep learning model underperforms compared to baselines	Model may not surpass traditional ML methods	Clearly document findings and contribute insight; maximise use of university resources to improve model tuning and evaluation
Computational constraints	Prolonged training times or restricted model size	Leverage University GPU clusters and free-tier HPC resources such as Google Colab Pro or Kaggle Kernels
Ethical concerns related to student data	Potential breaches of data usage policy or ethical standards	Confirm dataset licences and ethics approvals early; ensure anonymisation where applicable
Falling behind schedule	Risk of incomplete implementation or rushed final report	Follow Gantt milestones; reserve final weeks for catch-up if needed; adjust non-core goals accordingly
Scope creep	Attempting too many extensions beyond core AES system	Limit focus to clearly defined objectives; only pursue optional enhancements once core implementation is complete

Resources and Limitations

This project will draw upon a combination of computational, data, and academic resources available through the University of Bath and open-source platforms.

Resources

- **Datasets:** Publicly available Automated Essay Scoring datasets, particularly the ASAP (Automated Student Assessment Prize) dataset hosted on Kaggle. Other datasets may be used to supplement or test model generalisability, depending on availability.
- **Computational Resources:** Access to University of Bath's High-Performance Computing (HPC) infrastructure, including GPU support where possible. Cloud services such as Google Colab or Kaggle Kernels may also be used for experimentation or prototyping.
- **Libraries and Tools:** Python-based machine learning and NLP libraries, including Hugging Face Transformers, Scikit-learn, PyTorch or TensorFlow (depending on final implementation), and supporting libraries for data processing and visualisations.
- **Version Control and Documentation:** GitHub will be used for version control, issue tracking, and collaboration (should it be used anywhere). LaTeX will be used for academic report writing and formatting.
- **Academic Resources:** Access to online academic journals, papers, and resources via the University of Bath library (e.g., IEEE Xplore, ACL Anthology, JSTOR).

Limitations

- **Hardware Constraints:** Training large-scale transformer models on full-length essay data can be computationally expensive. Model size and training configuration may be constrained by available GPU access and memory limitations.
- **Dataset Quality and Scope:** The project relies on publicly available datasets, which may not reflect the full diversity of student writing, prompts, or demographic characteristics. Some datasets may lack metadata needed for fairness analysis.
- **Time Constraints:** The project is constrained by the fixed MSc dissertation timeline. As such, extensions such as feedback generation, multi-modal input (e.g., handwriting), or real-time deployment will only be explored if time permits.
- **Ethical and Legal Restrictions:** Use of student-generated content is limited to datasets with appropriate licenses or ethical clearance. No private or institutionally sensitive data will be used without explicit approval.
- **Model Interpretability Trade-offs:** Highly performant models may be less interpretable. There is a trade-off between achieving high predictive accuracy and ensuring the system remains explainable and educationally useful.