PROJECT ABSTRACT

Essay Scoring using BERT

A bert-base-uncased fine-tuning pipeline on the ASAP automated essay scoring benchmark (12,978 essays, 8 prompt sets) that predicts essay scores with per-trait feedback, evaluated with quadratic weighted kappa — delivered as a built-to-order student project with code, notebook, report, PPT and viva kit.

01. Title

Essay Scoring using BERT

02. Introduction / Background

Teachers spend hours grading essays and students wait weeks for feedback that arrives too late to act on. Automated essay scoring has been researched since the 1960s, but transformer models made it practical: BERT's contextual representations capture the coherence, vocabulary and argument structure that distinguish strong writing from weak. The ASAP benchmark, released through Kaggle by the Hewlett Foundation, provides 12,978 student essays across 8 prompt sets with human-resolved scores, and the field evaluates on it with quadratic weighted kappa (QWK). This project fine-tunes bert-base-uncased on ASAP with per-set regression heads and wraps it in a demo that scores essays with trait-level feedback.

03. Problem Statement

Manual essay grading is slow and inconsistent, while existing automated tools are either opaque or unreliable. The project addresses this by building a BERT-based scoring pipeline trained on the standard ASAP benchmark: it must handle 8 prompt sets with different score scales and genres, predict scores that agree with human raters as measured by QWK, and explain each score through trait-level feedback (content, organization, language, conventions) rather than a single inscrutable number.

04. Objectives

- Preprocess the ASAP dataset: WordPiece-tokenize essays (max 512 tokens), normalize scores per prompt set, and use the official evaluation splits.
- Fine-tune bert-base-uncased with a dropout + linear regression head per prompt set using AdamW (lr 2e-5, ~3 epochs).
- Evaluate with quadratic weighted kappa per set and compare against the human-human agreement reference (~0.75).

- Build a demo app with three views: the essay scorer with highlighted trait feedback, the ASAP dataset/rubric explorer, and the QWK evaluation view.
- Document the fairness limitation (US grade 7-10 training distribution; assistive, not authoritative scoring) and deliver the complete student kit.

05. Proposed Methodology

The methodology has four stages. (1) Data: ASAP essays are tokenized to 512 WordPiece tokens; scores are normalized per set because scales differ (Set 1: 2-12, Set 8: 0-60). (2) Model: bert-base-uncased's CLS embedding feeds a per-set dropout + linear regression head, trained with MSE on normalized scores using AdamW. (3) Evaluation: per-set QWK is computed on the official splits used in the ASAP literature; attention-adjacent token contributions are mapped back to highlighted spans for the trait feedback. (4) Inference: the demo tokenizes a pasted essay, runs the fine-tuned model, rescales to the set's range, and renders the score, trait bars and highlighted suggestions.

06. System Architecture / Working

The pipeline flows as: ASAP essays -> WordPiece tokenization (512) -> per-set score normalization -> bert-base-uncased fine-tuning with 8 regression heads -> per-set QWK evaluation -> exported weights. The demo app loads the weights for the interactive scorer: essay text in, predicted score plus four trait scores out, with highlighted spans carrying concrete feedback. The rubric view documents all 8 prompt sets; the evaluation view charts per-set QWK against the human-agreement reference. The notebook recomputes every metric the demo displays.

07. Hardware / Software Requirements

- Hardware: GPU recommended for fine-tuning (the notebook documents expected runtimes and a smaller-sample quick-run mode); CPU suffices for inference.
- Software: Python 3.10, PyTorch, Hugging Face transformers, scikit-learn, pandas, Matplotlib, Jupyter; a modern browser for the demo.
- Dataset: ASAP Automated Essay Scoring (Kaggle; download instructions in the setup guide).

08. Expected Outcomes

A fine-tuned BERT essay scorer achieving per-set QWK measured by the student's own notebook (design target: mean QWK approximately 0.75-0.80, near human-human agreement, stated as a target not a pre-claim); a demo that scores a sample persuasive essay with trait feedback; and a report documenting the per-set results, the error analysis (where long narratives and creative writing fail), and the fairness discussion.

09. Applications

Draft-feedback tools for classrooms, writing-practice apps giving instant trait-level feedback, placement-test triage assisted by human raters, and as a teaching example of transfer learning, regression on text, and rigorous NLP evaluation with QWK.

10. Future Scope

Trait-annotated training for finer feedback, cross-prompt generalization research, multilingual essay scoring with XLM-R, rubric-conditioned scoring where the teacher supplies the rubric, and longitudinal tracking of a student's writing progress across drafts.