Project Proposal: Automated IELTS Writing Evaluation System Using NLP

Trushar Ghanekar, Hoan Vu, Prachi Sheth

May 4, 2025

1 Introduction

IELTS Writing Task 2 [1] presents a significant challenge for test-takers who are seeking to demonstrate their English writing skills. The manual evaluation process for these essays can often be subjective and time-consuming, making it difficult to achieve a consistent grading. In addition, the essay question pool tends to be repetitive and restricted to a limited number of topics throughout an assessment cycle. This project aims to apply Natural Language Processing (NLP) [2] techniques to build a robust automated IELTS writing evaluation pipeline. Our main goals include collecting and cleaning datasets, developing a question generation model for the IELTS writing task 2, and creating an essay evaluation system.

1.1 Research Questions

- **RQ1:** How effective are transformer-based models [3] like BERT [4] compared to traditional models like logistic regression in predicting IELTS overall score?
- **RQ2:** Does incorporating semantic similarity (e.g., using BERT embeddings [4]) improve alignment between essays and prompts?

2 Methodology

2.1 Technical Approach

- **Data Preprocessing**
  - Remove unnecessary characters (e.g., punctuation, emojis).
  - Normalize text (e.g., lowercase, lemmatize, stop-word removal).
- **Feature Extraction**
  - Use TF-IDF [5] vectors for logistic regression.
  - Use BERT embeddings for transformer-based modeling.
- **Model Training**
  - Train a logistic regression model on extracted TF-IDF features.
  - Fine-tune a pre-trained BERT model using essay text and overall score.
- **Evaluation**
  - Use accuracy and F1 score to compare model performances.
  - Perform cross-validation and evaluate generalizability on the test set.

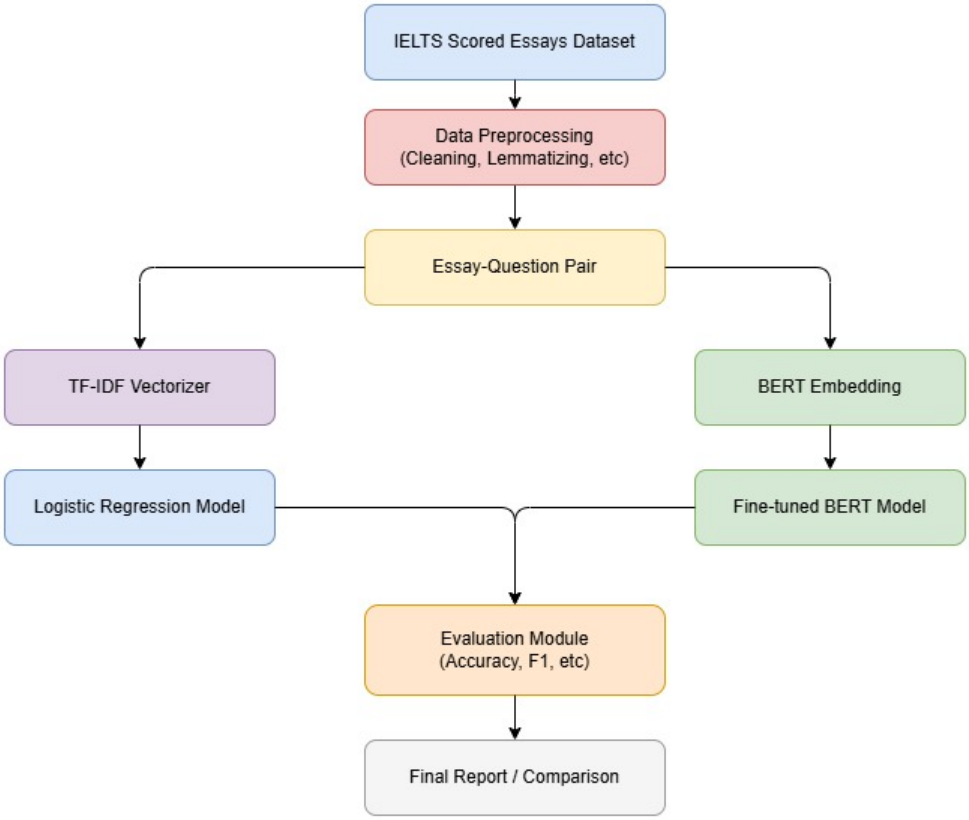

Figure 1: System Architecture of the Automated IELTS Writing Evaluation

2.2 NLP Techniques

- **Text Normalization:** Convert essays to lowercase, remove punctuation, and perform lemmatization to standardize the input text.
- **Tokenization:** Break down each essay into individual tokens (words or subwords) to enable further processing by models.
- **TF-IDF Vectorization:** Transform the textual data into numerical feature vectors that represent the importance of words across essays. We will use traditional model - logistic regression.
- **Transformer-Based Embeddings (BERT):** Leverage pre-trained transformer models to extract contextual embeddings from essay text. These embeddings capture semantic meaning and are used as input for deep learning models.
- **Vector Similarity Analysis:** Compute cosine similarity between essay embeddings and high-scoring reference responses to assess semantic alignment and coherence. This can serve as an auxiliary feature or evaluation metric.

3 Team Contributions

3.1 Shared Responsibilities

All Members: Data gathering and pre-processing, train/validation/test split creation, working process documenting and final project poster preparation.
Deliverables:

- A git repository with all materials to verify the experiments.
- A poster for the poster session.

3.2 Individual Responsibilities

3.2.1 Hoan Vu

Role: Implement and train the logistic regression model.

- Implement TF-IDF vectorization.
- Implement and train the logistic regression models (for both question generation and essay grading).
- Document model performances.

Deliverables: Trained logistic regression model.

3.2.2 Prachi Sheth

Role: Implement and train the BERT model.

- Generate BERT embeddings.
- Fine-tune a pre-trained BERT model on the essay dataset.
- Handle GPU-related setup if applicable and train the models (for both question generation and essay grading).
- Document model performances

Deliverables: Trained BERT model.

3.2.3 Trushar Ghaneka

Role: Develop the evaluation framework and calculate performance metrics.

- Design and implement evaluation framework (accuracy and F1 scores).
- Compute and compare accuracy and F1 scores for both models.
- Run cross-validation, write up results and insights.
- Semantic similarity analysis using cosine similarity on BERT embeddings

Deliverables: Detailed evaluation comparing both models.

4 Evaluation and Dataset

4.1 Dataset Description

For training and testing our system we are using these following datasets of publicly available essay(s):

- IELTS Writing Scored Essays Dataset on Kaggle

IELTS Writing Scored Essays Dataset [6] contains 1274 essay samples, each accompanied by various attributes such as the writing task question, human-written responses, and band scores across multiple categories (task response, coherence and cohesion, lexical resource, range accuracy, and overall score). The columns relevant to our analysis are listed below.

- **Task Type:** Essay task categories (either 1 and 2)
- **Question:** Questions or writing prompts to the essay
- **Essay:** Actual written essay responses submitted by the IELTS candidates
- **Overall:** Final scores assigned to each essay

Task Type	Question	Essay	Overall
2	Nowadays, not enough students choose science subjects in university in many countries. What are the reasons for this problem? What are the effects on society?	It is a serious problem that there is an inadequate number of students who tend to select science as their major subject in universities in many countries...	7

Table 1: IELTS Writing Scored Essays Dataset Example

4.2 Experimental Setup

We will evaluate the models on the IELTS Writing Scored Essays dataset using a train/validation/test split. The following metrics will be used to assess model performance:

- Accuracy - To assess overall correctness in band score prediction.
- F1 Score - To evaluate the balance between precision and recall, especially important if score distribution is imbalanced.

References

[1] IELTS Official Website, "Ielts writing task 1 and task 2." https://ielts.org/take-a-test/test-types/ielts-academic-test/ielts-academic-format-writing. Accessed: 2025-05-01.

[2] Wikipedia contributors, "Natural language processing." https://en.wikipedia.org/wiki/Natural_language_processing, 2025. Accessed: 2025-05-01.

[3] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2019.

[5] K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of Documentation*, vol. 28, no. 1, pp. 11–21, 1972.

[6] ibrahimmazlum , "Ielts writing scored essays dataset." https://www.kaggle.com/datasets/mazlumi/ielts-writing-scored-essays-dataset/data, 2023. Accessed: 2025-05-01.