UE Research Journal
Vol. 29 (2026): pp. 104-138

Automated Essay Evaluator Using Bidirectional Encoder Representations from Transformers Algorithm and Semantic Analysis

Kathleen M. Dimaano
College of Engineering
University of the East, Caloocan

Abstract

This study introduces the design and assessment of the Automated Essay Evaluator (AEE) system based on the use of the Bidirectional Encoder Representations from Transformers (BERT) algorithm and Semantic Analysis, which will help overcome some problems encountered with conventional essay evaluation methods. Using a developmental descriptive research design guided by the CRISP-DM process, the system integrates multiple Natural Language Processing models to assess grammar, structure, semantic relevance, and originality with a high degree of accuracy. Designed for use in educational institutions, the AEE integrates advanced Natural Language Processing (NLP) models such as BERT-base-uncasedd, all-MiniLM-L6-v2, spelling-correction-English-based, chatgpt-detector-roberta, and ms-macro-MiniLM-L6-v2 to assess grammar, structure, semantic relevance, and originality. The AEE system ensures wide-ranging evaluation because it determines if there are semantic relationships between essays, identifies grammatical errors, checks for plagiarism, and detects if the content is artificial intelligence (AI) generated. Statistical analysis revealed a strong linear association between the human consensus and automated scoring through Pearson's $r = 0.9700$ where $p < 0.001$. To verify fairness, a two-way ANOVA confirmed no statistically significant difference between human and automated scoring methods with value of $p = 0.297$, while the Intraclass Correlation Coefficient (ICC 2,1) yielded a value of 0.962, indicating excellent reliability. These results demonstrate that the application's Automated Scoring driven by semantic validation. The result demonstrates high accuracy when it comes to plagiarism and AI generated detection with a score of 94%. The ISO/IEC 25010 criteria for software quality factors were used, and they received very good ratings, especially in the Safety category with a mean weighted score of 4.86. The flexibility, adaptability, and usability aspects of the system were obtained through feedback from the experts who are teachers, students, and information technology practitioners.

Keywords: Automated Essay Scoring, Natural Language Processing, BERT Algorithm, semantic analysis, educational technology tools, essay evaluation, AI in education