

Automated Evaluation of Responses Using Embedding-Based Similarity with Explainable AI Support: A Comparative Study of all-MPNet-basev2 and all-MiniLM-L6-v2

Ravindrashankar M.^{1*}, Adeeba S.¹

¹Department of Computing and Information Systems,
Faculty of Computing, Sabaragamuwa University of Sri Lanka, Sri Lanka
*mravindrashankar@std.appsc.sab.ac.lk

With the rapid expansion of digital and large-scale learning environments, automated grading of student responses has gained greater significance among educators. In manual grading, there are many difficulties like time consuming, inconsistency, and difficulty to manage large scale responses. Advances in Natural Language Processing, especially transformer-based sentence embeddings, enables researchers an opportunity to develop more meaningful evaluation tools. Motivated by this need, this research examines the efficiency of semantic-similarity-based Auto-Grader system aided by Explainable AI (XAI) to provide a transparent and fair evaluation. This study focuses How well can transformer-based sentence embeddings evaluate essay and MCQ type responses? and to what extent can XAI make automatic grading judgments more transparent? The objectives include creating a semantic similarity based scoring mechanism, evaluating the performance of the all-MPNet-base-v2 and all-MiniLM-L6-v2 models, and integrating XAI techniques. The framework was tested on datasets of essay and MCQ type responses If the similarity of the response exceeding the predefined threshold, were classified correct. To determine the words which contributed mostly to the overall similarity, Local Interpretable Modelagnostic Explanations (LIME) and SHapley Additive exPlanations (SHAP) style approximation are used. Such explanations help instructors to know why an answer has been rated as right or wrong. Findings indicate evaluation results that all-MPNet-base-v2 scored 97.78% and 81.46% respectively, in MCQ answers and essays while all-MiniLM-L6-v2 scored 96.67% for MCQ and 74.26% for essay responses respectively. Overall allMPNet-base-v2 performed slightly well. In conclusion, this paper outlines a fair, scalable and interpretable automated grading system, suggesting adaptive feedback and future extensions of multimodal assessment.

Keywords: *Automatic Grading; Semantic Similarity; Sentence Embeddings; Explainable AI (XAI); Natural Language Processing*