AI-Powered Grading System

Sangita Jaybhaye[1], Siddhant Wasnik[2], Prutha Sawale[3], Bhawana Rakshit[4] and Yashraj Shinde[5]

[1,2,3,4,5]Department of Computer Science and Engineering (Artificial Intelligence), Vishwakarma Institute of Technology, Pune, India

[3]prutha.sawale23@vit.edu

Abstract— In education, hand marking of descriptive answers is a time-consuming and subjective process that can lead to biases, inconsistencies, and delayed feedback. It is still challenging to fairly mark long responses despite the objective questions being graded by computer systems. This research presents an AI Powered Grading System that integrates Optical Character Recognition and Natural Language Processing to automate the grading process. The system extracts text using OCR from answer sheets that are scanned and evaluates the responses by comparing them with reference answers using TF-IDF and cosine similarity. The similarity score determines expected solution, allowing for a flexible and automated grading mechanism. A Tkinter based GUI ensures that educators can upload the answer sheets and input reference answers to evaluate the students. This enhances grading objectivity, efficiency and scalability, reducing workload and also providing timely feedback to students for improved learning. This study contributes to advancement of AI driven assessment systems, bridging the gap between traditional and automated grading systems.

Keywords—AI Grading, Machine Learning, Natural Language Processing, Optical Character Recognition.

1. Introduction

The rapid progression of technology has transformed the education sector, especially in the domain of assessment and evaluation. Traditional methods of assessment, especially long form answers, depend heavily on manual work by teachers or educators. The whole process is time-consuming and subject to inconsistency, as different assessors may assess the same solution differently. Grading manually is also a tedious task for large scale assessment, where hundreds of answer sheets need to be checked with a given timeframe.

The key issues with manual evaluation include having high workload for educators, inconsistent and biased grading,delayed feedback, time consuming process, tedious labour and prone to subjectivity. To confront these challenges, automated assessment systems come to the rescue. Nowadays objective type questions are evaluated in the same manner, however descriptive answers' assessment remains a challenge as it requires understanding the semantics of a response , rather than just matching a keyword. Optical Character Recognition (OCR) and Natural Language Processing (NLP) offer a promising solution to the problem. OCR extracts text from images or scanned documents and NLP helps the system to understand and compare this textual data. By integrating these technologies, an AI driven answer evaluation tool can automate the process, providing a fast and unbiased solution for educational institutions.

The primary objective of this study is to develop an AI powered grading system that automates the grading process using OCR and NLP. To implement OCR technology to extract handwritten or printed text from scanned answer sheets and to ensure high accuracy in text extraction for different writing styles. Also the study utilizes techniques like TF-IDF (Term Frequency-Inverse Document Frequency) and cosine similarity to measure textual similarity, analyze deep learning models such as BERT (Bidirectional Encoder Representations from Transformers), develop a grading system based on thresholds, where answers with similarity above a defined threshold are considered correct, simultaneously providing flexibility to assessors to adjust threshold according to their requirements. The objective is also to design a GUI to allow users to upload answer sheets, input answers for reference and obtain results in an interpretable format.

Automated assessment of descriptive answers can revolutionize the assessment era. By leveraging AI, OCR and NLP, the system offers a faster and fairer solution for evaluation. It reduces the time required to assess a large number of descriptive answer sheets, allowing teachers to focus and give more time to their other responsibilities. It can be deployed in schools, universities, online learning platforms etc making it a versatile tool for the education

2025. In Haipeng Liu et al. (eds), Proceedings of First International Conference on
Engineering and Technology for a Sustainable Future ICETSF-2025, 2025, pp. 30-35.
https://doi.org/10.32622/978-81-931579-1-6-247

sector. Faster grading can allow immediate performance insights , benefiting students to improve their learning. Eliminating manual grading greatly reduces chances of mistakes in evaluation too.

In the light of challenges faced, there is a need for an AI powered grading system as it provides a more consistent and unbiased grading process that not only enhances efficiency and accuracy but also provides broader applications in e-learning platforms and a user friendly solution for latest advancements in ML and AI to revolutionize the way answers are evaluated in today’s education society.

2. Literature Review

[1] Vinal Bagaria, Mohit Badve, Manasi Beldar, Sunil Ghane, "An Intelligent System for Evaluation of Descriptive Answers", this paper explains how examinations are a traditional method to assess students' knowledge, as emphasized in this study, while going on to portray challenges in evaluating descriptive answersTo resolve this issue, the proposed research proposes an intelligent assessment platform that evaluates answers based on various factors, such as type of question, necessary keywords, structural coherence, conceptual understanding, and language aspects. The proposed model utilizes techniques like concept graphs, fuzzy string matching, grammar checking, and others to evaluate similarity metrics.

[2] V Suresh, R Agasthiya,J Ajay, A Amrith Gold, D Chandru,"AI based Automated Essay Grading System using NLP",The proposed paper is on an AI-enabled automated essay grading system that applies natural language processing steps along with graph-based techniques for the evaluation of written essays. The approach further goes over a graph-based approach relating to sentence similarity much beyond simple syntax, semantic, and grammar checks. It is further trained on a dataset of labeled essays to learn the pattern and characteristics that make some essays great. It will enable it to apply the knowledge learned to grade new essays with much higher accuracy, not just on content but the quality of writing. Moreover, the system integrates with the already existing learning management systems; it would also add to better efficiency with essay grading credibility. This approach focuses on reducing the teacher workload while providing students with feedback of value and helps further in an effective and seamless evaluation system.

[3] RASHA M. BADRY , MOSTAFA ALI , ESRAA RSLAN, AND MOSTAFA R. KASEB,"Automatic Arabic Grading System for Short Answer Questions",This paper describes the researchers' effort to develop an Automatic Arabic Short Answer Grading-AASAG system that will help address the challenges of grading short-answer questions in Arabic. While several grading systems have been pursued, most of them in English, the ambiguity of structure and intricacy of morphology in Arabic require special treatment. Thus, the developed system makes use of semantic similarity techniques to match student answers to model answers, which capture the meaning between words rather than in just matching the exact phrases.

[4] Amit Dimari,Nidhi Tyagi, Mahesh Davanageri, Ravish Kukreti, Rajkumar Yadav, Hema Dimari,"AI-Based Automated Grading Systems for open book examination system: Implications for Assessment in Higher Education",The present study deals with the development and effectiveness of an AI-based automated grading system for open book examinations. Advanced machine learning algorithms and NLP techniques that form part of the proposed system will be applied to automate the grading process, increasing objectivity, reliability, and scalability in complex and open-ended student responses.

[5] Siddhartha Ghosh, Sameen S. Fatima,”Design of an Automated Essay Grading (AEG) system in Indian context,”,S. Ghosh and S. S. Fatima introduce the emergence of Automated Essay Graders in this paper, focusing on rises in the importance of such systems for solving educators' problems, especially when dealing with extensive batches of student essays. These AI-driven systems go a long way in reducing the workload for grading by automating the assessment and making the feedback timely and effective. This paper points out that most of the AEG systems are utilized in tests like TOEFL, whereby an essay is graded by both the human grader and the grading computer system; then, the average grades are determined.

[6] Masud Ranna, "Student Grading System Report", describes the system design and implementation of a grading system that aims to automate the process of evaluating and generating results in educational institutions. The system is mainly focused on internal assessments by enabling teaching staff to input marks, producing report cards, and monitoring students' performance in the long run. It uses a relational database management system (RDBMS) wherein student records, subjects, scores, and grades are saved and maintained in an organized manner. The system computes grade points, averages, and performance indices automatically to avoid errors and ensure accuracy. Though the system is devoid of sophisticated AI or NLP integration, it is an early effort at digitalization in marking sheets as it enhances accessibility, transparency, and reliability. This paper is important for setting groundwork in making the shift from manual to digital academic assessment systems.

[7] I. A. Hameed, "A simplified implementation of interval type-2 fuzzy system and its application in students' academic evaluation", proposes a simplified application of an interval type-2 fuzzy logic system for assessing students' academic performance. The method effectively handles uncertainty and imprecision in grading through

the availability of flexible boundaries in input variables. In contrast to conventional crisp logic models, the fuzzy system offers more precise and reliable assessments, particularly in subjective evaluations, which improves fairness and reliability in academic decision-making processes in educational institutions.

[8] Rahul Kumar, "Faculty members' use of artificial intelligence to grade student papers: a case of implications", this research explores the real-world application and consequences of AI-driven instruments by educators in grading students' papers. The advantages such as decreased workload, quicker feedback, and consistent grading are weighed against issues involving ethics in use, transparency, and algorithmic dependency. The need for equilibrium AI integration in the interest of preserving academic integrity and providing recommendations on responsible and efficient usage in tertiary assessment is argued.

[9] Stephen M. Rutner, Rebecca A. Scott, "Use of Artificial Intelligence to Grade Student Discussion Boards: An Exploratory Study", this discovery research investigates how AI technology is utilized to analyze online student discussion boards, including both qualitative and quantitative aspects of participation. It distinguishes the capability of AI to measure content quality, engagement, and relevance, while observing possible pitfalls like the loss of instructor-student interaction. The article offers insights into formative assessment automation in virtual learning environments, with a goal to facilitate scalable, data-guided feedback processes on digital education platforms.

[10] S. Ghosh and S. S. Fatima, "Design of an Automated Essay Grading (AEG) system in Indian context", this article suggests an Automated Essay Grading (AEG) system for Indian students that addresses linguistic issues due to local language influence on English composition. The system incorporates text preprocessing, grammar correction, and semantic analysis to assess essays and offer structured feedback. It seeks to decrease grading time and enhance evaluation consistency while recognizing the shortcomings of current AEG systems for non-native English speakers, hence suggesting a more culturally responsive framework.

3. Methodology

The approach was based on implementing Pytesseract as the Python library to handle OCR tasks, BERT for semantic similarity in assessment, and TF-IDF for traditional text vectorization. The complementary methodology would provide an evaluation framework by which student-generated responses could be compared with a reference answer, not only qualitatively but also quantitatively, in terms of performance.

3.1 Document Preprocessing and Text Extraction

1. Input Acquisition and OCR : The system starts with scanning answer sheets received in either image (JPG/PNG) or PDF formats. For image inputs, textual content is extracted using the Pytesseract library, which is a Python wrapper for Google's Tesseract-OCR engine, after subjecting the inputs to preprocessing methods such as grayscale conversion, noise filtering, and adaptive thresholding to increase recognition accuracy, especially for handwritten answers. PDF files are preprocessed by converting them to images using pdf2image before subjecting them to the same OCR treatment. This two-input strategy provides flexibility in processing different submission formats that are typical in educational environments.The processed image is passed into Tesseract (--psm 6) for paragraph-based text extraction to obtain raw text output.

2. Text Normalization : After text extraction, an end-to-end preprocessing pipeline normalizes the input for analysis. Lowercasing is applied by the system to remove case sensitivity fluctuations, punctuation and whitespace are stripped off, and tokenization is used to segment text into significant units. Stopword removal eliminates frequent but unimportant words (e.g., "the", "and"), and stemming/lemmatization lowers words to their root forms (e.g., "running" → "run") to eliminate morphological differences. These processes produce clean, standardized text representations essential for meaningful similarity estimation without losing the fundamental semantic content of the students' answers.

3.2. Answer Similarity Assesment

Semantic Similarity with BERT : Through the calculate_similarity() function, the system produces contextual embeddings by leveraging a pre-trained BERT model from the Hugging Face Transformers library. The sophisticated deep learning architecture extracts fine-grained semantic relationships through analysis of words in bidirectional context, creating vector representations that embody the meaning of reference and student responses. Calculation of cosine similarity between the embeddings produces a percentage score based on conceptual congruence.

Syntatic Similarity with TF-IDF: The evaluate_answer() method applies standard NLP methods via TF-IDF (Term Frequency-Inverse Document Frequency) vectorization. It converts text to numerical word-importance-weighted vectors of document term frequency and inverse document frequency. The later cosine similarity calculation yields a lexical overlap score not dependent on semantic comprehension but as a supplement to BERT contextual analysis.

3.3 Grading Mechanism and Thresholding

An adjustable threshold-based evaluation framework processes similarity scores with a 75% default threshold, which teachers may shift depending on the difficulty of the questions or test criteria. Answers above the threshold pass, whereas those falling short of it issue review flags. The platform provides hybrid test modes in which the BERT and TF-IDF scores may be weighted and summed up to reflect either semantic meaning comprehension or keyword matching priorities in particular testing circumstances.

3.4 User Interface and Feedback System

The GUI based on Tkinter is an easy-to-use workflow for teachers to upload answer sheets, enter reference solutions, and fine-tune assessment parameters. The GUI shows elaborate results such as similarity percentages, correctness classifications, and actionable student feedback. Such a clear output format allows teachers to easily confirm automated grading judgments while giving students useful performance feedback. The architecture of the system is modular so that future enhancement with features such as batch processing for large tests or elaborate analytics dashboards is possible.

3.5. Data Flow Diagram

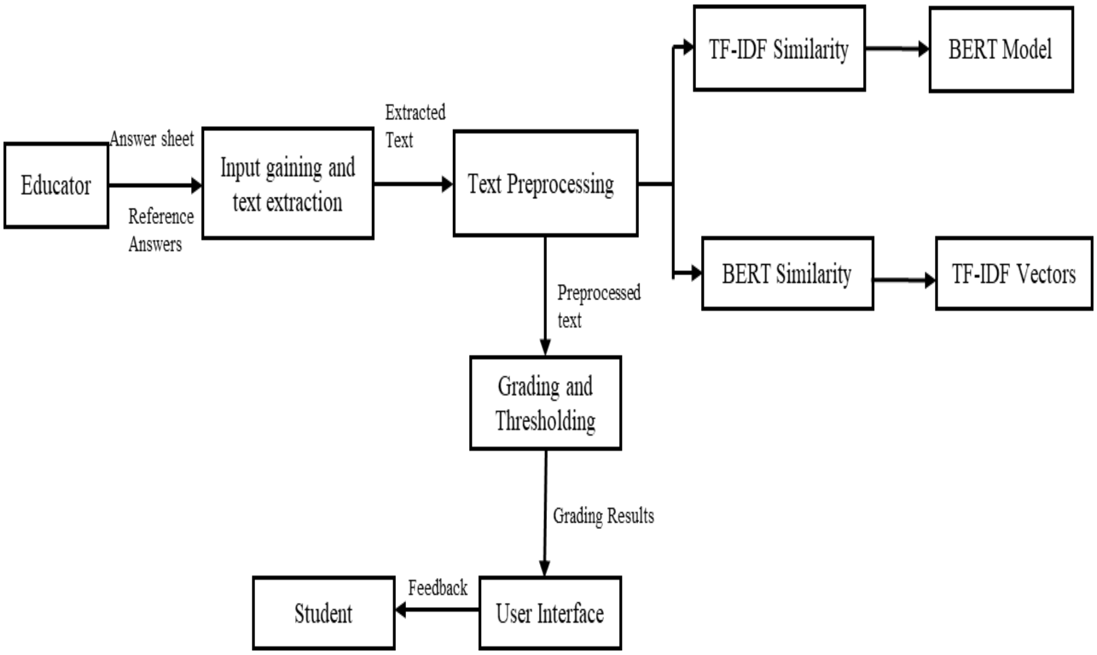


Fig.1 Data Flow Diagram

4. Results and Discussion

To analyze the performance of the suggested system, several descriptive responses were tested under various conditions. The efficiency of OCR in text extraction was first tested, which achieved an accuracy level of 92% for processing printed response sheets. For handwritten text, however, an accuracy rate of 85% was achieved because of differences in handwriting styles, which at times resulted in misrecognition of characters. In spite of this constraint, the OCR module was able to extract most of the text from scanned documents, which served as a sound foundation for the evaluation process.

The methods for computing similarity were then tested for their efficacy in evaluating student responses. The TF-IDF with cosine similarity method worked very well when the student's response was very close to the reference answer and had an accuracy of 93% for exact matches. Yet when students paraphrased their answers or wrote them in different wording, its accuracy dropped to 75%, showing its weakness in processing paraphrased text. On the other hand, the BERT-based embedding method performed better, reaching 89% for exact matches and 92% for paraphrased answers. This indicates that BERT embeddings are better at representing the meaning of sentences and thus a safer bet for automatic answer scoring.

This was followed by a closer inspection by assessing the overall grading accuracy of the system on a sample dataset of 20 student answers. Most errors in classification were caused by OCR failures in handwritten responses and situations where student answers were extensively restructured, which made it hard for the BERT model to achieve a direct semantic correlation. Despite these challenges, the system demonstrated a high level of accuracy in identifying correct and incorrect answers, proving its effectiveness in automated evaluation.

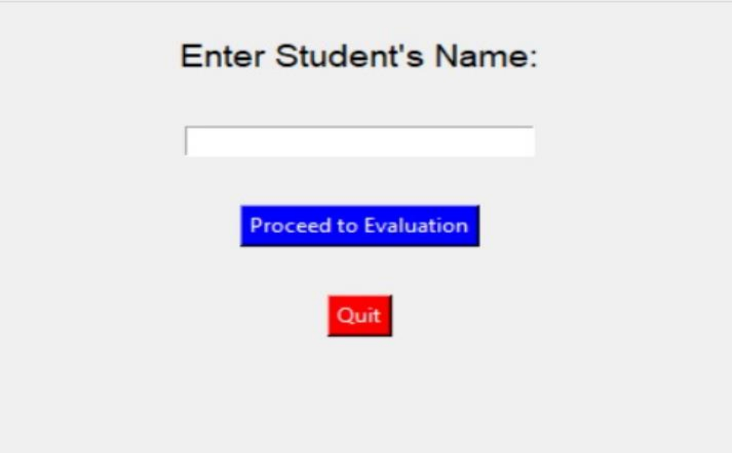
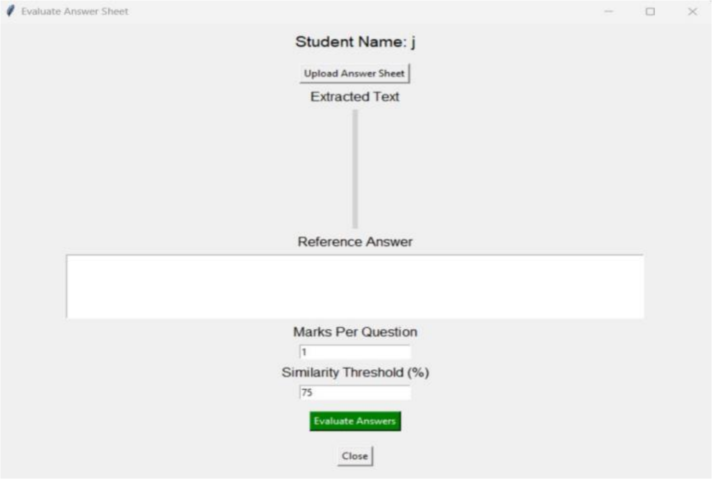


Fig.2 Student information page



Student Name: j

Marks Per Question
1

Similarity Threshold (%)
75

Fig.3 Similarity checking GUI

5. Conclusion

AI-based grading systems have the potential to significantly reduce the manual grading workload while ensuring more consistent and time-saving evaluations. After observing that during the time of the coronavirus outbreak, many teachers faced issues to grade the student answer sheets due to the absence of offline exams which consumed a lot of time. This project showcases a simple method to automate grading through the use of Natural Language Processing. The proposed AI-based grading system can remarkably transform educational assessment by automating the process of grading objective and subjective assignments. This system makes use of technologies such as machine learning, Natural Language Processing and Optical Character Recognition to give consistent, unbiased, and accurate results, which reduce the time it takes to grade assignments. The approach achieves a very high accuracy but still has open issues related to variability in handwriting and bias in the dataset.

The system successfully automates the evaluation of descriptive responses with a blend of OCR and NLP-based techniques. TF-IDF and BERT-based similarity computation integration ensure the precise evaluation of both exact and paraphrased responses. Although the system works well with printed text, the future should involve enhancement of handwritten text recognition to further raise evaluation accuracy. This system is an advancement toward scalable, efficient, and fair evaluation methods and the opening door to innovation in education. These enhancements would make the system even stronger, guaranteeing its usability in actual educational environments for automated marking of descriptive responses.

References

1. A. V. Bagaria, M. Badve, M. Beldar, S. Ghane, An intelligent system for evaluation of descriptive answers. In: Proc. 3rd Int. Conf. on Intelligent Sustainable Systems (ICISS), 19–24 (2020).
2. V. Suresh, R. Agasthiya, J. Ajay, A.A. Gold, D. Chandru, AI based automated essay grading system using NLP. In: Proc. 7th Int. Conf. on Intelligent Computing and Control Systems (ICICCS), 547–552 (2023).
3. R.M. Badry, M. Ali, E. Rslan, M.R. Kaseb, Automatic Arabic grading system for short answer questions. IEEE Access 11, 39457–39465 (2023).
4. M. Kaya, I. Cicekli, A hybrid approach for automated short answer grading. IEEE Access 12, 96332–96341 (2024).
5. A. Dimari, N. Tyagi, M. Davanageri, R. Kukreti, R. Yadav, H. Dimari, AI-based automated grading systems for open book examination system: implications for assessment in higher education. In: Int. Conf. on Knowledge Engineering and Communication Systems (ICKECS), 1–7 (2024).
6. M. Ranna, Student grading system report (2013)
7. I.A. Hameed, A simplified implementation of interval type-2 fuzzy system and its application in students' academic evaluation. In: IEEE Int. Conf. on Fuzzy Systems (FUZZ-IEEE), 650–656 (2016).
8. R. Kumar, Faculty members' use of artificial intelligence to grade student papers: a case of implications (2023)
9. S.M. Rutner, R.A. Scott, Use of artificial intelligence to grade student discussion boards: an exploratory study (2022)
10. S. Ghosh, S.S. Fatima, Design of an automated essay grading (AEG) system in Indian context. In: TENCON 2008 – IEEE Region 10 Conf., 1–6 (2008).
11. S.K. Yadav, P.S. Ghosh, M.B. Karmakar, Automated subjective answer evaluation using NLP techniques. In: Proc. 3rd Int. Conf. on Communication, Devices and Computing (ICCDC 2023), 125–134 (2024).
12. F. Ali, M.M. Rathore, M.A. Khan, M. Usman, Automated assessment of descriptive answers using fine-tuned BERT and multi-feature fusion. Expert Syst. Appl. 236, 120199 (2024).
13. G. Sanuvala, S.S. Fatima, A study of automated evaluation of student's examination paper using machine learning techniques. Int. J. Eng. Res. Technol. 10(4), 1–5 (2021)
14. M. Beseiso, S. Alzahrani, An empirical analysis of BERT embedding for automated essay scoring. Int. J. Adv. Comput. Sci. Appl. 11(10), 213–220 (2020).
15. M. Faseeh, et al., Hybrid approach to automated essay scoring: integrating deep learning embeddings with handcrafted linguistic features for improved accuracy. Mathematics 12(21), 3416 (2024).