International Research Journal of Education and Technology

Peer Reviewed Journal, ISSN 2581-7795

Automating Handwritten Answer Evaluation: A Deep Learning and OCR Integrated Approach

Pof.A.H.Pawar[1], Vaishnavi Chavan[2], Vedantika Pol[3] , Harshad Shinde* [4] , Sonyabapu Thorat* [5]

[1]Professor, Department Of Information Technology, SVPM's College Of Engineering, Malegaon (Bk), Maharashtra, India.

[2,3,4,5]Student, Department Of Information Technology, SVPM's College Of Engineering , Malegaon (Bk), Maharashtra, India.

Abstract - Subjective answer evaluation remains a complex and time-consuming task in education. This paper presents an automated evaluation system that uses advanced Natural Language Processing (NLP) techniques to assess student answers against teacher-provided reference solutions. The system employs pre-trained BERT Transformers from Hugging Face's library to encode both student and reference answers into semantic vectors, with cosine similarity used to measure semantic closeness. A rule-based scoring mechanism assigns scores based on defined similarity thresholds.

The system supports the evaluation of optional (OR) questions by calculating scores for multiple responses and selecting the highest similarity value. A user-friendly front-end is developed using the Streamlit framework, enabling teachers to manage subjects, classes, and upload PDF answers. PyMuPDF (fitz) is used for answer extraction, with the data stored in an SQLite database for processing. The system normalizes per-question scores and aggregates them for comprehensive exam evaluation.

Experimental results show that the system achieves a high correlation with human graders, outperforming traditional keyword-based methods. This work contributes to the development of scalable, efficient, and accurate grading tools, reducing manual effort and supporting personalized feedback for learners.

Key Words: Automated Evaluation, Handwritten Answer Grading, PaddleOCR, BERT Transformer, Cosine Similarity, NLP Models, Educational Assessment, Machine Learning, Multi-Language Support

1.INTRODUCTION

The rapid advancement of technology has significantly transformed various sectors, including education. In particular, the assessment process within educational institutions has evolved, moving from traditional manual grading methods to more automated and efficient systems. This paper introduces an automated answer evaluation system designed to streamline the grading process, ensuring consistency, efficiency, and scalability.

Traditional manual grading is labor-intensive and prone to inconsistencies, especially when evaluating subjective responses. With the increasing volume of assessments in educational institutions, there is a pressing need for automated systems that can efficiently and accurately evaluate student responses. Natural Language Processing (NLP), a subfield of artificial intelligence, has emerged as a pivotal technology in automating the evaluation of textual data. By enabling machines to understand, interpret, and generate human language, NLP facilitates the development of systems that can assess the quality of student responses based on semantic content rather than mere keyword matching.

This paper presents the design and implementation of an automated answer evaluation system that leverages Sentence Transformers to assess student responses. The system computes cosine similarity scores between student answers and reference answer keys to determine the degree of similarity. Based on predefined thresholds, the system assigns marks to student responses, ensuring an objective and consistent evaluation process.

Furthermore, the integration of a database management system allows for the storage and retrieval of evaluation results, supporting functionalities such as teacher registration, class management, and result tracking. This holistic approach not only streamlines the grading process but also provides valuable insights into student performance, aiding educators in identifying areas that require attention.

The subsequent sections of this paper delve into the system architecture, methodology, implementation details, and evaluation results, demonstrating the efficacy and potential of automated answer evaluation in modern educational settings.

2. Related Work

In recent years, there has been significant progress in automating the evaluation of subjective answers through Natural Language Processing (NLP) and Machine Learning (ML). Early methods, such as those by Mittal and Devi (2016), combined Latent Semantic Analysis (LSA) and BLEU metrics with fuzzy logic to evaluate both syntactic and semantic similarities in student responses. Their approach provided a foundation for integrating semantic analysis with automated scoring.

George and Rexie (2013) explored the role of Information Extraction (IE) in subjective answer evaluation, emphasizing techniques such as part-of-speech tagging and sentence parsing. They highlighted the importance of NLP in understanding user queries and extracting relevant data, paving the way for more sophisticated methods that focus on answer quality and relevance.

In more recent work, Girkar et al. (2021) combined grammatical correction, keyword extraction, and semantic similarity measures using tools like NLTK WordNet, aiming for a more nuanced evaluation that accounts for context and variations in phrasing. Their approach demonstrated improved accuracy, achieving a 90.3% success rate in differentiating varying answer qualities.

Kudale et al. (2023) and Kumari et al. (2023) further advanced this domain by incorporating deep learning models and advanced NLP techniques. Kudale et al. (2023) used Cosine and Semantic Similarity measures, while Kumari et al. (2023) integrated BERT embeddings, FuzzyWuzzy for keyword matching, and grammar APIs to improve grading consistency and accuracy. The use of pre-trained models such as BERT allows for richer semantic understanding and context preservation.

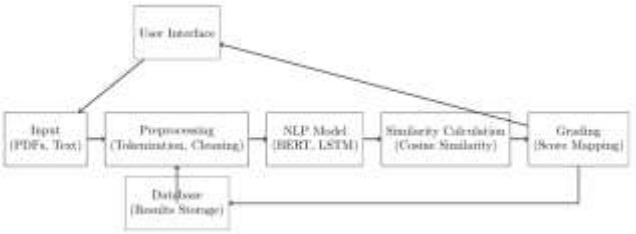
Recent research also explored deep learning architectures like Siamese LSTM networks. Lokhande et al. (2022) implemented Manhattan LSTM (MaLSTM) and its variants to measure semantic similarity and achieve real-time grading accuracy. Their results indicated that multi-layered models significantly improved performance, suggesting that advanced deep learning techniques could play a key role in subjective answer evaluation.

Building upon these advancements, our approach leverages a **Bidirectional LSTM (BiLSTM)** architecture for better context capture and sequence modelling.The **attention mechanism** is integrated to focus on the most relevant parts of the answers, further improving the model's ability to distinguish between key aspects of student responses. Additionally, **BERT** embeddings are utilized for capturing deep semantic meaning and ensuring that even responses with varied phrasing are evaluated with high accuracy.

Our system aims to address the limitations of previous methods, such as poor handling of context and inconsistencies in manual grading, by combining the power of **deep learning models** and **semantic analysis**. We demonstrate that our approach not only enhances grading accuracy but also ensures fairness and scalability, capable of handling diverse student answers with high efficiency

3. Proposed System

The proposed system architecture consists of multiple components working together to achieve the goal of evaluating student answers. The architecture follows a modular approach, ensuring flexibility, scalability, and efficiency in evaluating subjective answers.



Proposed System Architecture Diagram

- Submodule 1: Preprocessing This submodule is responsible for cleaning and structuring the input text data from the teacher's answer key and the student's response. It includes:

Subjective Answer Evaluation–

Text Cleaning: Removing unnecessary characters, punctuations, and noise. Document Parsing: Converting input documents (like PDFs) to structured

Text.– Tokenization: Splitting text into smaller tokens, such as words or phrases.

Text Embedding: Transforming text into vector representations for further analysis.

- Submodule 2: NLP Model Integration This submodule handles Natural Lan gauge Processing tasks by integrating pre-trained models like BERT or LSTM. It extracts semantic and contextual information from the text, which is essential for evaluating the answers.

- Submodule 3: Similarity Computation In this part, the system uses Cosine Similarity to measure how closely the student's answer matches the model answer based on their vectorized representations.

1.1System Overview

The proposed system, **SubjectiveAnswerEvaluator**, is designed to automate the evaluation of subjective answers provided by students in educational assessments. The system aims to improve the accuracy, consistency, and efficiency of grading subjective answers by leveraging Natural Language Processing (NLP) techniques, machine learning models, and similarity measurement algorithms. The primary objective is to assist educators by reducing manual grading time while providing real-time, constructive feedback to students.

Developed an end-to-end AI-driven system for extracting, correcting, and evaluating handwritten student answers from scanned PDFs.

The solution automatically processes uploaded handwritten answer sheets by performing Optical Character Recognition (OCR) using PaddleOCR, followed by multi-level text corrections including spelling correction (SpellChecker), grammar enhancement (LanguageTool), and context improvement using a transformer-based BART model.

Cleaned and structured answers (Answer 1, Answer 2, etc.) are saved into organized PDFs and text files. These extracted answers are further evaluated against teacher-provided answer keys using advanced Natural Language

Processing (NLP) similarity techniques to calculate student scores automatically.

The entire system supports multiple users (teachers and students), enables easy answer management, and provides downloadable reports and scorecards.

1.2Methodology

The proposed system, leverages a combination of advanced techniques in **Natural Language Processing (NLP)**, **Similarity Measurement**, **Machine Learning**, and **Feedback Generation** to automate the evaluation process of subjective answers. This study proposes a systematic approach for extracting, correcting, and structuring handwritten textual answers from scanned PDF documents. The methodology consists of multiple phases including pre-processing, optical character recognition, text correction, and final output generation. Each phase is detailed as follows:

The following key techniques are employed:

1. Preprocessing

A comprehensive multi-stage pipeline to convert handwritten PDF documents into structured, machine-readable text. Initially, each page of the input PDF is converted into high-resolution images (300 DPI) to ensure that handwriting details are preserved, enhancing the subsequent text extraction process. Deep learning-based Optical Character Recognition (OCR) is then applied to detect text regions, handle rotation variations, and accurately recognize handwritten and printed content. To address the inevitable inaccuracies introduced during OCR, a multi-level text correction pipeline is employed, consisting of word-level spell correction, sentence-level grammar enhancement, and context-aware rewriting using a transformer-based language model, thereby ensuring grammatical accuracy, semantic coherence, and overall readability. Post-correction, the text is segmented into distinct answer units based on predefined markers, enabling structured extraction of responses. Finally, the processed content is serialized into two output formats: a professionally formatted PDF document and a plain-text file, ensuring logical organization, human readability, and compatibility with downstream applications.

2. Natural Language Processing (NLP)

The system utilizes state-of-the-art NLP models, specifically **BERT** and **Sentence-BERT**, to capture the contextual meaning of the student's response. These models are trained to generate high-quality text embeddings that represent the semantic meaning of the answer. The embeddings are critical for accurately measuring the relevance and coherence of student responses in comparison to the reference answer. By encoding both the student's and reference answers into vector representations, these models ensure that semantic similarity is effectively captured.

3. Cosine Similarity

To compare the student's answer with the reference answer, **Cosine Similarity** is employed. This metric measures the cosine of the angle between the two embeddings, providing a numerical representation of their similarity. The closer the cosine value is to 1, the more similar the student's response is to the reference answer. This method allows for precise comparison and evaluation based on the contextual meaning of the responses rather than exact word matching.

4. Machine Learning Models

These categories are based on various features such as **answer length**, **keyword usage**, and **semantic content**. The machine learning models are trained to learn the patterns from the answers and assign them to appropriate categories, helping to enhance the accuracy and consistency of the evaluation process.

5. Feedback Generation

Based on the results of the similarity measurement and classification, the system is designed to automatically generate **personalized feedback** for the student. The feedback is aimed at guiding students on how to improve their responses, focusing on aspects such as clarity, relevance, and completeness. By providing automated suggestions, the system enhances the learning process by giving students immediate insights into the quality of their answers and how they can improve them for future assessments.

4. CONCLUSIONS

This study presents an integrated framework for automating the extraction and evaluation of handwritten student responses, leveraging advancements in Optical Character Recognition (OCR), Natural Language Processing (NLP), and machine learning. The proposed system utilizes PaddleOCR for accurate handwriting recognition, followed by advanced text correction processes, including spelling and grammar adjustments, and semantic refinement through the BERT transformer model. This ensures the extracted text is syntactically accurate and semantically coherent.

The evaluation module employs BERT Transformers to generate semantic embeddings and applies cosine similarity measures to compare student submissions with reference answers. Additionally, the system supports dynamic handling of various question types, threshold-based scoring, and comprehensive grade calculation. A secure and intuitive teacher interface facilitates the management of student data, answer key uploads, multi-student evaluations, and result storage.

By automating the grading process, this framework offers significant improvements over traditional manual evaluation. It enhances grading consistency, reduces human effort, and provides a scalable solution that can be adapted to diverse educational contexts. Future work will focus on expanding multilingual support, adapting to different handwriting styles, and incorporating advanced feedback generation mechanisms to offer a more personalized learning experience.

While the current system is effective, there remain several opportunities for further enhancement. Expanding its multilingual capabilities would broaden its applicability across regions with diverse linguistic needs. Incorporating handwriting style adaptation techniques would improve accuracy for different writing styles

REFERENCES

[1]Bashir, M. F., Arshad, H., Javed, A. R. (Member, IEEE), Kryvinska, N., & Band, S. S. (Senior Member, IEEE). (2021). Subjective Answers Evaluation Using Machine Learning and Natural Language Processing. *IEEE Access*, doi:10.1109/ACCESS.2021.3130902. Received November 3, 2021; accepted November 22, 2021; date of publication November 25, 2021; date of current version December 7, 2021.

[2] S. Lokhande, U. Chaudhary, A. Singh, P. Gaikwad, H. Guleria, and S. Pawar, "Automated Subjective Answer Evaluation System," *Journal of Algebraic Statistics*, vol. 13, no. 3, pp. 1108–1113, 2022. [Online]. Available: https://publishoa.com.

[3] P. Patil, S. Patil, V. Miniyar, and A. Bandal, "Subjective Answer Evaluation Using Machine Learning," *International Journal of Pure and Applied Mathematics*, vol. 118, no. 24, pp. [Page Numbers], May 2018. [Online]. Available: http://www.acadpubl.eu/hub/`

[4] H. Mittal and M. S. Devi, Subjective evaluation: A comparison of several statistical techniques, Appl. Artif. Intell., vol. 32, no. 1, pp. 8595, Jan. 2018.

[5] G. Kudale, N. Mali, N. Suryawanshi, M. Bansode, and R. Agarwal, "Automated Subjective Answer Evaluation Using NLP," *International Journal of Creative Research Thoughts (IJCRT)*, vol. 11, no. 5, May 2023. [Online]. Available: http://www.ijcrt.org

[6] A.H. Pawar, V. Chavan, V. Pol, S. Thorat, and H. Shinde, "Review on subjective answer evaluation," *International Research Journal of Modernization in Engineering Technology and Science*, vol. 6, no. 10, Oct. 2024. [Online]. Available: https://www.doi.org/10.56726/IRJMETS62554