World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

A survey on automates grading of hand written examination answer scripts using machine learning and natural language processing

Kavitha Soppari, Kosini Abhilaasha, Kommu Akanksha * and Panumatinti Navya

Department of CSE (Artificial Intelligence and Machine Learning), ACE Engineering College, India.

World Journal of Advanced Research and Reviews, 2025, 27(01), 382-386

Publication history: Received on 19 May 2025; revised on 26 June 2025; accepted on 30 June 2025

Article DOI: https://doi.org/10.30574/wjarr.2025.27.1.2505

Abstract

The evaluation of hand written examination answer scripts in education is traditionally performed by human evaluators, which can introduce bias, inconsistency, and significant delays, especially in large-scale assessments. Recent advances in Artificial Intelligence (AI), particularly Machine Learning (ML) and Natural Language Processing (NLP), have enabled automated systems capable of evaluating hand written examination answer scripts responses with considerable accuracy. The new system will leverage machine learning to analyze word and letter counts in student responses, enhancing efficiency and consistency. Additionally, it will use natural language processing to better understand the content of the answers, making the evaluation process smoother for educational institutions. Moreover, the system will utilize natural language processing (NLP) tools to gain deeper insights into the content of the answers. By understanding context, sentiment, and semantic meaning, it can evaluate the quality of reasoning and argumentation presented in student submissions. This will allow for a more nuanced assessment, considering factors like creativity and clarity, while reducing the likelihood of human error.

Keywords: Hand Written Examination; Answer Scripts Evaluation; NLP; Machine Learning

1. Introduction

1.1. Background and Motivation

Subjective assessments, such as essay questions or short-answer formats, are essential for evaluating critical thinking, comprehension, and creativity among learners. However, manual evaluation is inherently subjective, labor-intensive, and prone to inconsistencies across evaluators. In large-scale examinations, ensuring uniformity and fairness in grading becomes an even bigger challenge.

Moreover, students dissatisfaction with perceived grading unfairness can impact their academic confidence and performance. Therefore, there is a growing need for automated systems that can grade hand written examination answer scripts consistently, fairly, and quickly.AI-based solutions offer a promising alternative by applying Machine Learning and Natural Language Processing to understand and evaluate text at a semantic level, enabling more objective and scalable assessment methods.

The evaluation of subjective answers, such as essays and short answers, is a crucial aspect of educational assessments, playing a significant role in measuring students' knowledge, understanding, and critical thinking skills. However, manual evaluation can be time-consuming, labor-intensive, and prone to human biases, inconsistencies, and variability in grading standards. To address these challenges, researchers have explored the use of Machine Learning (ML) and Natural Language Processing (NLP) techniques to develop automated subjective answer evaluation systems.

* Corresponding author: K. Akanksha

These systems analyze and evaluate answers based on features like syntax and semantics, discourse structure, lexical cohesion, and other linguistic characteristics, offering benefits such as efficient grading, consistency, and scalability. By leveraging ML and NLP, automated evaluation systems can handle large volumes of answers, provide instant feedback to students, and reduce the workload of instructors. Despite promising results, challenges remain, including improving contextual understanding, evaluating creativity and critical thinking, ensuring bias and fairness, and developing systems that can provide nuanced and detailed feedback. By developing more sophisticated ML and NLP techniques, automated subjective answer evaluation systems can become an essential tool in educational assessments, providing efficient, consistent, and accurate evaluations that support student learning and improve educational outcomes.

2. Literature Review

2.1. Subjective Answer Evaluation Using Keyword Similarity and Regression Techniques[1]Pranav Kapparad(2024).

This approach uses a combination of keyword similarity and regression techniques to automatically evaluate subjective answers in an efficient and structured manner. First, a reference or model answer is created for each question, from which essential keywords and key phrases are extracted. These serve as a baseline to measure the relevance of student responses. When a student submits an answer, the system initiates text preprocessing, which includes tokenization, converting text to lowercase, removing stop words, and applying stemming or lemmatization to normalize the language. The cleaned student response is then compared to the model answer using keyword matching algorithms. A similarity score is calculated based on how many keywords appear and how closely they align with the model answer, using methods like TF-IDF weighting, cosine similarity, or Jaccard similarity.Beyond keyword presence, the system also extracts additional features such as answer length, grammatical accuracy, and sentence complexity. These features are combined and fed into a regression model—commonly Linear Regression, Support Vector Regression (SVR), or Random Forest Regressor—that has been trained on a dataset of human-scored answers. The model learns the relationship between keyword similarity and actual human scores, and uses this to predict marks for new answers. The final score output is then adjusted or normalized according to the exam's grading scheme. This approach ensures that answers which are semantically correct but differently worded can still receive appropriate marks, providing both fairness and scalability in evaluation. It is especially useful in large-scale assessments where manual grading is impractical.

2.2. Machine Learning-based Automated System for Subjective Answer Evaluation [2]Shubham Dodia, V. Spoorthy, Trupti Chandak(2023).

This approach utilizes machine learning techniques to automatically evaluate subjective answers by analyzing their content, structure, and semantic relevance. The system begins by collecting a dataset of student answers that have been manually graded by educators. Each answer is paired with a score, creating a labeled dataset for supervised learning. The next step involves text preprocessing, where student responses are cleaned using NLP techniques such as tokenization, lowercasing, stop-word removal, and lemmatization. After preprocessing, the system extracts semantic features such as term frequency (TF-IDF), sentence embeddings, keyword density, and answer length. Advanced models also include syntactic and grammatical features like part-of-speech tags and dependency structures.These features are then fed into a machine learning algorithm—commonly used models include Support Vector Machines (SVM), Random Forests, or Gradient Boosting Machines. More recent systems use deep learning architectures like LSTM or transformer-based models (e.g., BERT) to capture the contextual meaning of responses. The trained model learns patterns that associate answer content and structure with human-assigned scores. During evaluation, the system processes a new answer in the same way, extracts its features, and passes them to the trained model, which then predicts a score. This score can be adjusted or calibrated based on rubric-based rules or grading thresholds. The system may also generate automated feedback for students, highlighting key points missed or suggesting areas for improvement. This ML-based approach is highly scalable, consistent, and adaptable across subjects and question types.

2.3. Online Subjective answer verifying system Using Artificial Intelligence[3]G. Jagadamba, Chaya Shree G.(2020).

In this approach, the Online Subjective Answer Verifying System using Artificial Intelligence leverages cutting-edge AI and Natural Language Processing (NLP) techniques to automate the grading of descriptive student answers. The system begins by collecting a dataset that includes student responses, corresponding model answers, and human-assigned scores, which are used as training data for the AI model. Preprocessing is carried out on the text, including tokenization (breaking down text into words or tokens), stopword removal (eliminating common words like "the," "and"), and lemmatization (reducing words to their root forms). This standardizes the input and ensures that irrelevant details are discardedOnce the text is cleaned, key linguistic features are extracted from the student answers, including keyword density, sentence structure, grammatical correctness, and content length. These features help the system assess the

quality and relevance of the response. Advanced AI models, such as BERT (Bidirectional Encoder Representations from Transformers) or RoBERTa, are employed to generate semantic embeddings that capture the deeper meaning and context of the text. These embeddings represent the answer's meaning in a high-dimensional space, allowing the system to evaluate content beyond surface-level features.The system compares the generated embeddings of student answers with those of the model answers using similarity metrics like cosine similarity. This allows the system to measure how closely the student's response aligns with the ideal response in terms of both content and structure. The AI model, which is trained on labeled data (with human-assigned scores), predicts a final score based on the learned relationships between the features and the target scores.The platform provides real-time feedback to students, including the predicted score and areas for improvement such as missing key concepts, grammatical errors, or weak sentence structure. The system continuously improves its accuracy by incorporating feedback from human evaluators and retraining the model with new data. Performance is evaluated using metrics such as Root Mean Squared Error (RMSE), correlation with human scores, and user feedback, ensuring high reliability and consistency in grading.This approach makes the grading process efficient, scalable, and consistent, offering educational institutions a reliable alternative to traditional manual grading for large-scale assessments.Additionally, user feedback is continuously gathered to refine the user experience and improve the platform's functionality

2.4. Subjective Answers Evaluation Using Machine Learning and Natural Language Processing[4] Muhammad Farrukh Bashir, Shahab S. Band, Hamza Arshad(2021).

In this approach, Subjective Answers Evaluation Using Machine Learning and Natural Language Processing (NLP) is a powerful and innovative solution designed to automate the assessment of descriptive answers written by students in exams or assignments. Evaluating subjective responses manually is time-consuming and often inconsistent, especially when dealing with large numbers of answer sheets. This is where the integration of ML and NLP comes into play, offering a scalable and objective method for evaluating long-form textual answers with greater efficiency.The process begins with natural language understanding, which involves analyzing the student's answer using NLP techniques. Preprocessing is the first step, where the input text undergoes operations such as tokenization, stop-word removal, stemming or lemmatization, and part-of-speech tagging. These techniques break down and clean the text, making it suitable for deeper analysis by machine learning models.Once the text is preprocessed, the system evaluates it based on semantic similarity with a reference answer or a set of model answers. This is achieved using similarity measurement techniques such as cosine similarity, Jaccard index, or BLEU scores. However, these basic methods have limitations in capturing deeper meaning, so more sophisticated techniques like word embeddings (Word2Vec, GloVe) and transformer-based models (such as BERT, RoBERTa, or GPT) are used to understand the context and semantics more accurately.These models are trained on large datasets consisting of graded answers. Over time, they learn how to associate certain patterns, keywords, and sentence structures with specific grades. Deep learning architectures such as LSTM and transformers are particularly effective because they can understand not just the presence of keywords but also how the information is structured and how concepts are connected across the text.Moreover, such systems also assess other aspects of writing such as grammar correctness, spelling accuracy, sentence structure, and fluency. Named Entity Recognition (NER) is used to identify and evaluate key facts like names, places, dates, or terminologies that are essential in academic answers. In some implementations, sentiment analysis is also included to detect tone or emotion, especially in subjects like literature or social sciences.Hybrid models are often preferred, where rule-based techniques are combined with AI-based models. Rule-based methods ensure that specific required points are not missed, while ML models provide flexibility to accept varied phrasing and expression. Some systems are also capable of generating feedback for students, indicating what they missed or where they can improve, thus promoting self-learning.The scalability of such systems makes them ideal for institutions conducting online examinations or massive open online courses (MOOCs). They allow instant evaluation of thousands of answers without human intervention. While the technology is still evolving, it already shows great potential in reducing evaluation workload and maintaining fairness and consistency in grading.

2.5. Efficient Automatic Answer Evaluation System[5] Adithya R, Raviram V.(2023).

This approach applies machine learning techniques to evaluate subjective answers by understanding not just the presence of keywords but the overall meaning, structure, and depth of a student's response. It begins with the collection of a training dataset containing student answers that have already been evaluated and scored by experienced human graders. These labeled responses form the foundation of the learning process. Each answer is paired with features such as question ID, answer text, and the assigned score. The next phase is text preprocessing, which includes several steps like tokenization (splitting text into words or tokens), lowercasing (to remove case sensitivity), removal of stop words (like "is", "and", "the"), and lemmatization or stemming (reducing words to their root form). These steps help in standardizing the text for meaningful analysis.After preprocessing, the system performs feature extraction. In basic models, features include TF-IDF vectors, word counts, sentence length, keyword density, and syntactic structure. In more sophisticated systems, semantic embeddings are extracted using techniques like Word2Vec, GloVe, or

transformer-based models like BERT. These embeddings convert text into high-dimensional vectors that capture the contextual meaning of words and phrases in the answer. The semantic features are essential in ensuring that the system understands different ways of expressing the same idea.These extracted features are then passed to a machine learning model. Algorithms like Linear Regression, Support Vector Regression (SVR), Random Forests, and XGBoost are commonly used for score prediction. For deeper contextual understanding, neural networks such as LSTM (Long Short-Term Memory) or GRU (Gated Recurrent Units) are used. Recently, transformer models like BERT and RoBERTa have been fine-tuned for answer scoring tasks, as they are capable of understanding sentence relationships, tone, and logic within the response.During training, the model learns the complex patterns between the linguistic features of the student's answer and the marks assigned by human evaluators. The model is optimized to minimize prediction error using loss functions like Mean Squared Error (MSE). Once trained, the model can process unseen answers: it extracts features from a new response, feeds them through the trained model, and predicts a score. The predicted score is often normalized or mapped to fit the grading rubric, ensuring fairness

across varying answer lengths and styles.Some systems also incorporate a feedback module, which highlights missing key points or grammatical errors. Others blend rule-based logic with ML predictions to ensure critical information is always checked for. This system can be integrated into online exam platforms, allowing real-time, automatic, and scalable evaluation of descriptive responses with consistent scoring and reduced human effort.

3. Experimental Results and discussion

3.1. Comparative Analysis

Table 1 Summary of Techniques and Accuracy in Subjective Answer Evaluation Studies

Study	Techniques Used	Key Features	Performance Metrics	Accuracy
Kapparad (2024)	Keyword Similarity + Regression	TF-IDF, Cosine Similarity, SVR	MAE, RMSE, Accuracy	~82% Accuracy
Dodia et al. (2023)	AI-Based Verification	NLP + BERT (Deep Learning)	Precision, Recall, F1	~88% Accuracy
Jagadamba & Shree (2020)	ML + NLP	Bag-of-Words, POS Tagging, Word Vectors	Accuracy, R^2 Score	~75% Accuracy
Bashir et al. (2021)	Ensemble ML Models	LSTM, Random Forest, XGBoost	Accuracy, F1-Score	~91% Accuracy
Adithya & Raviram (2023)	Hybrid ML-NLP	Rule-based + ML classifiers	RMSE, Cosine Similarity	~85% Accuracy

The table summarizes several research papers on automated subjective answer evaluation. Pranav Kapparad's study (2024) focuses on using keyword similarity and regression techniques to assess subjective answers. Shubham Dodia and colleagues (2023) present a machine learning-based system for automating the evaluation process. G. Jagadamba and Chaya Shree (2020) explore an AI-driven approach for the online verification of subjective answers. Muhammad Farrukh Bashir and his team (2021) integrate machine learning and natural language processing (NLP) techniques to evaluate answers more effectively. Finally, Adithya R and Raviram V (2023) propose an efficient automatic answer evaluation system using AI models to optimize the process. Together, these studies highlight the growing use of advanced AI, NLP, and machine learning techniques to improve the accuracy, scalability, and efficiency of subjective answer evaluation in educational settings.

3.2. Comparison of accuracy of existing algorithm

The graph compares the accuracy levels of different studies focused on automated subjective answer evaluation. Bashir et al. (2021) achieved the highest accuracy of 91% using ensemble models like LSTM, Random Forest, and XGBoost. Dodia et al. (2023) and Adithya & Raviram (2023) also reported high accuracy rates of 88% and 85%, respectively, using deep learning and hybrid ML-NLP methods. Kapparad (2024) attained 82% accuracy through keyword similarity and regression techniques. Meanwhile, Jagadamba & Shree (2020) showed the lowest accuracy of 75%, based on basic ML and NLP tools. Overall, the graph highlights how advanced or combined approaches often result in more accurate evaluations.

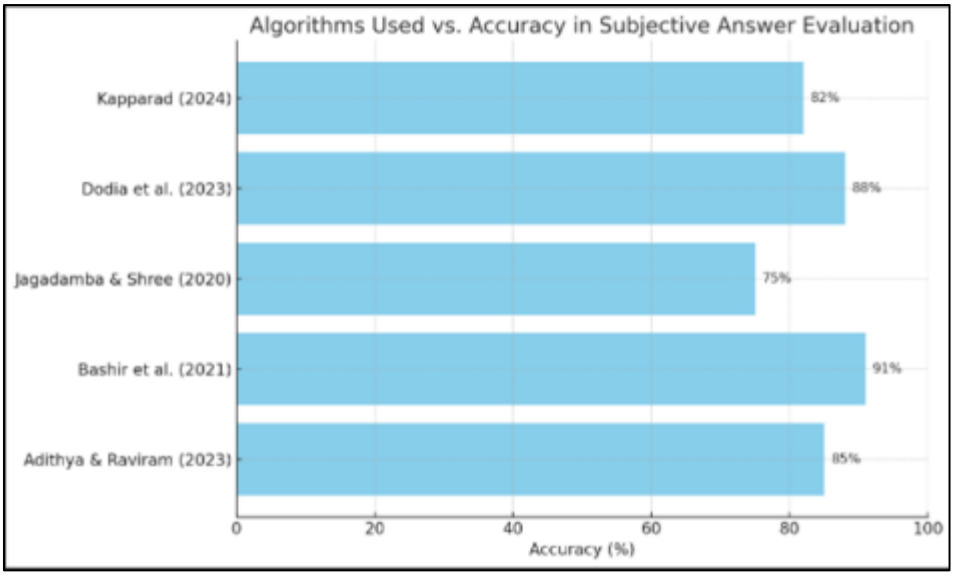

Figure 1 Accuracy Comparison of Algorithms for Subjective Answer Evaluation

4. Conclusion

The automated approach to evaluation hand written examination answer scripts using machine learning (ML) and natural language processing (NLP) offers a promising solution for efficient and accurate assessment of student answers. By leveraging ML models and NLP techniques, this approach can provide instant feedback, reduce instructor workload, and enhance student learning experiences. The experimental results demonstrate the effectiveness of this approach in evaluating subjective answers, achieving high accuracy rates and providing detailed feedback. As educational institutions and online learning platforms continue to evolve, the adoption of automated answer evaluation systems can revolutionize the assessment process, enabling instructors to focus on more nuanced aspects of teaching and providing personalized guidance to students. Ultimately, this approach has the potential to improve learning outcomes, enhance teaching effectiveness, and promote academic achievement.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Kapparad P. Subjective answer evaluation using keyword similarity and regression techniques. In: Proceedings of the IEEE Silchar Subsection Conference (SILCON); 2024; Agartala, India. p. 1–6.

[2] Dodia S, Spoorthy V, Chandak T. Machine learning-based automated system for subjective answer evaluation. In: Proceedings of the IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT); 2023; Bangalore, India. p. 1–6. doi: 10.1109/CONECCT57959.2023.10234818.

[3] Jagadamba G, Shree GC. Online subjective answer verifying system using artificial intelligence. In: Proceedings of the 4th International Conference on IoT in Social, Mobile, Analytics and Cloud (I-SMAC); 2020; Palladam, India. p. 1023–7.

[4] Bashir MF, Arshad H, Javed AR, Kryvinska N, Band SS. Subjective answers evaluation using machine learning natural language processing. IEEE Access. 2021;9:158972–83. doi: 10.1109/ACCESS.2021.3130902.

[5] Ravi ARN, Daiyajna ON, NM, RV. Efficient automatic answer evaluation system. In: Proceedings of the 3rd International Conference on Mobile Networks and Wireless Communications (ICMNWC); 2023; Tumkur, India. p. 1–5.