Subjective Answer Evaluation using Machine Learning and Natural Language Processing

Sejal Jagtap[1], Vaishnavi Joshi[2], Sanika Nikam[3], Nanidi Randive[4], Prof. M.A.Gade[5]
Information Technology JSCOE Pune, India

***Abstract*—This paper presents an Automatic Subjective Answer Evaluation (ASAE) system that automates the grading of subjective answers using machine learning and natural language processing (NLP)The evaluation of subjective answers plays a crucial role in the teaching and learning process. With the growing need for efficient and accurate grading systems, automatic evaluation of answers has become essential. However, existing systems often yield mediocre results, especially when evaluating short or long subjective answers. Traditional methods focus on keyword matching between the student's response and a reference answer, but these systems fail to deliver optimal results. Short answers, with their limited number of keywords, require special attention, particularly in calculating the weighting score. This study aims to evaluate the performance of existing frameworks for automatic grading of long and descriptive answers and suggests improvements for better accuracy and consistency. By analyzing the current mechanisms, this research seeks to enhance the overall effectiveness of automated answer evaluation systems.**

Keywords*—*Automated Essay Evaluation, Automatic Scoring, Automatic Grading, String Similarity, Content-Based Similarity, Natural-Language Processing

I. INTRODUCTION

In the modern education system, assessments play a pivotal role in measuring students' understanding and academic progress. While traditional assessment methods, such as multiple-choice questions and true/false tests, are widely used due to their convenience, they fall short when it comes to evaluating complex learning outcomes, including critical thinking, problem-solving abilities, and the synthesis of ideas. Subjective answers, which require students to explain concepts in their own words, are better suited to assess these advanced skills. However, evaluating these responses manually is both labor-intensive and prone to inconsistencies, even when grading rubrics are applied.

Despite the introduction of automated grading systems, many of these remain inadequate in assessing subjective answers with the accuracy and fairness that manual grading offers. Current systems generally rely on simple keyword matching or pattern recognition, which works well for short responses but struggles with longer, more complex answers. Short answers, in particular, present a challenge as they contain fewer keywords, making it harder to capture the full context and meaning of the response. As a result, the existing automated grading systems often fail to provide accurate feedback or comprehensive scores for subjective answers.

This paper addresses these challenges by proposing an Automatic Subjective Answer Evaluation (ASAE) system that incorporates machine learning and natural language processing (NLP) techniques to evaluate answers in a more context-aware manner. The proposed system goes beyond keyword matching by analyzing the semantic meaning, coherence, and structure of the responses, which is particularly crucial for evaluating long and descriptive answers. By focusing on the underlying meaning rather than just surface-level keywords, the ASAE system aims to enhance grading accuracy and reliability.

While the demand for automated grading systems has grown in response to this challenge, existing solutions for evaluating subjective answers have shown limited success. Most systems still rely on basic keyword matching or pattern recognition techniques, which are often ineffective in assessing long, descriptive answers. These systems struggle to account for the underlying meaning, context, and coherence of the responses, especially when dealing with answers that contain a limited number of keywords, such as short answers. Furthermore, maintaining scoring consistency and ensuring fairness in a subjective evaluation process remains an unresolved issue. As noted in existing research, the process of grading subjective answers manually not only requires significant time and effort but also presents challenges in achieving consistent and accurate results, even with grading rubrics in place.

Another key aspect of this research is the exploration of privacy and security concerns related to online grading systems. As education increasingly shifts to digital platforms, ensuring the privacy of student data becomes paramount. In line with recent developments in privacy-preserving technologies, this study considers the use of federated learning to protect students' personal information during the evaluation process. Federated learning enables the system to train models locally on users' devices without transmitting sensitive data to central servers, providing a secure way to evaluate responses while maintaining privacy. This aligns with modern requirements for ensuring that sensitive data is not misused, as seen in other applications such as online learning monitoring systems .

The major contributions of this paper are as follows:

- Development of an automated system that evaluates subjective answers by understanding their underlying meaning, context, and structure.
- Incorporation of privacy-preserving techniques like federated learning, ensuring that student data remains secure throughout the grading process.
- Improved accuracy and consistency in the evaluation of short and long descriptive answers compared to traditional keyword-based methods.
- A scalable and efficient solution that can be integrated into educational platforms to streamline the grading process and reduce teacher workload.

The structure of the paper is outlined as follows: Section II provides a review of the existing challenges in subjective answer evaluation and highlights relevant research. Section III describes the methodology used to develop the ASAE system, including the machine learning and NLP models employed. Section IV presents the experimental results, showcasing the effectiveness of the proposed system. Finally, Section V concludes the paper and discusses future directions for enhancing automated answer evaluation systems.

II. RELATED WORK

In recent years, there has been growing interest in developing automated systems for evaluating subjective answers, especially in educational settings. With advancements in natural language processing (NLP), machine learning (ML), and deep learning (DL), researchers have made significant strides in creating systems that can automatically assess and grade open-ended student responses. These systems aim to streamline grading, reduce human bias, and provide timely feedback to students. However, despite these advancements, several challenges remain, including accurately interpreting diverse writing styles, handling ambiguous or incomplete answers, and ensuring fairness and transparency in grading. Moreover, issues related to data privacy, especially in the context of student assessments, and the computational complexity of deep learning models are also significant barriers to the widespread adoption of these technologies in educational environments.

The paper [1] Kapoor et al. conducted an in-depth analysis of automated answer evaluation systems using machine learning techniques. Their research focused on evaluating subjective answers, both short and long, where traditional manual grading poses significant challenges due to its time-consuming nature and inconsistency in scoring. The study reviewed various existing systems and approaches that utilize string similarity, content-based similarity, and Natural Language Processing (NLP) to enhance the accuracy of automated grading. Methods such as cosine similarity, n-grams, and Latent Semantic Analysis (LSA) were identified as popular techniques in automated subjective answer grading. The authors traced the evolution of this field, starting from Ellis Page’s pioneering work in 1966 on computer-based grading, to modern advances in Automated Essay Scoring (AES) and Automated Essay Evaluation (AEE) systems. Despite the progress, they noted several difficulties faced in this domain. Key challenges include achieving high accuracy and consistency comparable to human graders, especially for short answers with limited keywords. Semantic understanding of descriptive answers remains problematic, as it requires advanced NLP models capable of grasping context and logical structure. Additionally, computational complexity and the high cost of developing these systems hinder their widespread adoption. Human bias in manual grading also complicates the creation of unbiased training data, further affecting the reliability of automated systems. The authors concluded that while current systems show promise, significant improvements are necessary to overcome these challenges and develop robust, scalable solutions for automated subjective answer evaluation.

Meanwhile, in [2], Mahalakshmi et al. focused on addressing the complexities of evaluating subjective

answers using a combination of machine learning and natural language processing (NLP) techniques. Their study emphasizes the inefficiency and time-consuming nature of manual grading, which is often affected by factors such as fatigue and bias. The authors proposed an automated system architecture that incorporates GloVe word embedding, cosine similarity, and Word-Cloud to enhance the accuracy of grading. Their approach involves extensive data preprocessing, including text cleaning, case folding, and special character removal, to prepare raw answers for evaluation. The system uses GloVe to capture semantic relationships between words, enabling more meaningful comparisons between student responses and reference answers. Cosine similarity is employed to compute the closeness between answers, providing a quantitative similarity score that serves as the basis for grading. The study also conducted a rigorous experimental evaluation to demonstrate the system's effectiveness. Despite these advancements, the authors acknowledge several challenges. These include handling the wide range of vocabulary and synonym usage in subjective answers, ensuring context sensitivity, and managing computational overhead for large datasets. Their work contributes to the ongoing development of efficient, scalable solutions for automated subjective answer evaluation, aiming to improve the consistency, fairness, and speed of academic assessments.

According to [3] the paper The proposed approach focuses on addressing the inefficiencies and inconsistencies associated with manual grading by utilizing techniques such as data preprocessing, feature extraction, and classification. During preprocessing, unnecessary content like headers and footers is removed, and the textual data is transformed into tokens or vectors for analysis. Feature extraction identifies key aspects of the answers, such as relevance, coherence, organization, and grammar. These features are then used as input for a classification model. The system employs the Multinomial Naive Bayes (MNB) algorithm, a probabilistic method that uses training data to categorize answers based on extracted features. The methodology is designed to ensure objectivity and efficiency, providing a scalable solution for subjective answer evaluation. The paper highlights several issues encountered in the process of automating subjective answer evaluation. One significant challenge is the lack of high-quality, labeled datasets required to train machine learning models effectively. Additionally, the variability in answer formats and the difficulty in capturing semantic nuances, such as creativity and coherence, pose substantial obstacles.

In contrast, the paper in [4] The paper presents a system for automatic evaluation of descriptive answers using natural language processing (NLP) and machine learning (ML). The implementation involves extracting text from answer scripts, generating summaries using keyword-based summarization, and computing multiple similarity measures like cosine similarity, Jaccard similarity, bigram similarity, and synonym similarity. These measures are used to evaluate answers against reference solutions. The pre-processing steps include tokenization, stop-word removal, lemmatization, and bigram creation, followed by feature extraction for scoring. The similarity measures are assigned weights, determined through surveys, to calculate the final score for each answer.

However, the system faced several challenges. Key issues included the manual assignment of weight values to similarity measures, which could introduce bias, and the slight discrepancies observed between automated and manually assigned scores in some cases. Additionally, while the approach demonstrated efficiency, the reliance on predefined parameters and lack of advanced learning mechanisms limited its adaptability. The authors proposed addressing these challenges in future work by integrating a machine learning model to automate weight determination and exploring more effective summarization techniques for improved scoring accuracy.

According to [5] The paper proposes a novel system for evaluating subjective answers using machine learning (ML) and natural language processing (NLP). The approach combines traditional NLP techniques like tokenization, lemmatization, and stemming with advanced similarity metrics such as Word Mover's Distance (WMD) and Cosine Similarity, alongside classification algorithms like Multinomial Naive Bayes (MNB). A corpus of subjective questions and answers was created, with annotators identifying key elements such as keywords and essential context. Answers are compared against predefined solutions using similarity metrics to score responses based on relevance and context. The scores are then refined by training a machine learning model, which improves over time and eventually serves as a standalone evaluator.

The primary challenges faced include the absence of a publicly available, high-quality dataset for training and testing, reliance on manual annotation for data preparation, and the limitations of existing similarity measures like TF-IDF, which often lose semantic context. Despite these obstacles, the proposed system achieved high accuracy (up to 88%) and demonstrated the potential for scalability and domain-specific improvements with further training and optimization. Future work aims to refine the word2vec model for domain-specific applications and enhance dataset quality to improve performance further.

In summary, various research efforts have tackled the problem of automatic grading for subjective answers, highlighting the challenges of context understanding, scalability, and fairness. The proposed system in the context of your project builds upon these insights by focusing on improving the accuracy and fairness of automatic subjective answer evaluation. By employing advanced NLP techniques, machine learning models, and incorporating feedback loops for continuous improvement, your system aims to address many of the limitations identified in previous works. Additionally, ethical considerations, such as ensuring data privacy and transparency, are integral to the design of your system.

III. SYSTEM ARCHITECTURE AND DESIGN

Step-by-step description of the working process for Automatic Subjective Answer Evaluation System based on system architecture.

A. Web Application Framework:

This system is built using Django as the web application framework to create a user-friendly interface and manage backend functionalities. Django provides a structured approach to collecting user input and rendering responses. And A form-based interface allows students to enter their subjective answers. Django's forms module is used for creating input fields. For Example: A text area field where students input their answers.

B. System Functioning Steps:

Step 1: User Input through Django Interface:

A form in Django collects textual responses from users. Example: A text area for students to input their answers. Django's form-handling capabilities are leveraged to validate and process inputs.

Step 2: Preprocessing the Input (NLP Step):

Prepare input data for analysis. Splitting text into words or sentences. Excluding common, irrelevant words. Converting words to their root form.

Step 3: Feature Extraction

Convert textual data into a machine-readable format for NLP models. Word embeddings or vectorization transforms textual data into vectors.

Step 4: Model Evaluation using NLP

Measure the similarity between student responses and model answers. Used Techniques Cosine Similarity it Computes similarity by measuring the cosine of the angle between two text vectors. Also used Semantic Similarity Uses context-based matching for deeper understanding.

Step 5: Data Storage and Management

Store user inputs, model answers, and similarity scores. Django's ORM interacts with a relational database. User data and evaluation results are stored using models.

C. Rule-Based Algorithms

1. BERT (Bidirectional Encoder Representations from Transformers)

BERT is a deep learning-based language model that understands the context of words within a sentence. Unlike traditional models that read text in one direction (left-to-right or right-to-left), BERT processes words in both directions simultaneously, enabling it to grasp the full context of a word. In this project, BERT helps analyze student answers by understanding their semantic meaning and comparing them with reference answers, ensuring context-aware evaluation.

2. Cosine Similarity

Cosine similarity measures the similarity between two texts by treating them as vectors in a multidimensional space. It calculates the cosine of the angle between these vectors, resulting in a similarity score between 0 (completely different) and 1 (completely similar). This technique is used to evaluate how closely the student’s answer matches the model answer without requiring exact word matches, making it effective for subjective evaluations.

3. Word2Vec

Word2Vec is a technique that represents words as numerical vectors based on their contextual usage. Words appearing in similar contexts have similar vector representations. In this project, Word2Vec allows the system to understand synonyms and related

phrases, enhancing the flexibility and semantic understanding during answer comparison.

4. WordNet

WordNet is a lexical database for the English language, containing relationships between words such as synonyms, antonyms, and hierarchies. It helps the system understand different word meanings and relationships, contributing to more accurate semantic analysis when comparing student answers to reference answers.

5. Word Mover's Distance (WMD)

Word Mover's Distance calculates the minimum distance needed to transform one text into another by moving words. Using Word2Vec representations, WMD provides a sophisticated measure of similarity by considering both the meanings and order of words in student and reference answers, allowing for better handling of flexible phrasing.

6. Multinomial Naive Bayes

Multinomial Naive Bayes is a probabilistic classification algorithm commonly used in text classification tasks. It works by applying the Bayes theorem, using word frequencies from training data to classify answers as correct or incorrect. This method is efficient and suitable for determining answer correctness based on word occurrence probabilities.

This combination of algorithms and techniques enables your system to evaluate subjective answers effectively by incorporating semantic similarity, contextual understanding, and classification strategies.

By combining rule-based approaches and NLP techniques, this system provides a robust evaluation mechanism for subjective answers.

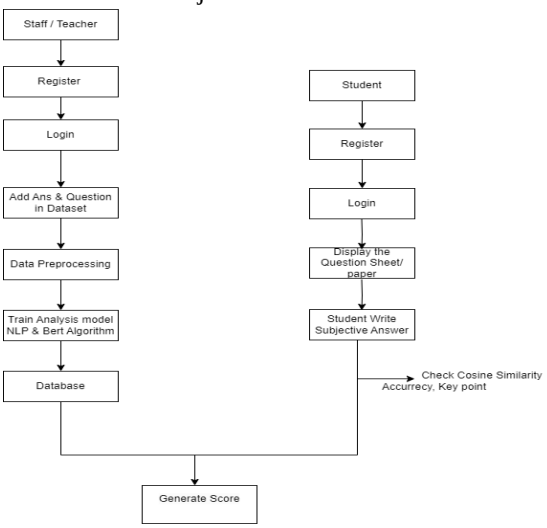

Fig 1. System Architecture

IV. RESULT AND ANALYSIS

A. Evaluation Metrics

This section presents the outcomes of the proposed Automatic Subjective Answer Evaluation system and analyzes its performance based on various metrics, such as accuracy and efficiency. The results are discussed to highlight the effectiveness of the system in automatically evaluating subjective responses provided by students.

1) Accuracy: Accuracy measures the proportion of correct predictions out of the total number of predictions. It is calculated using the formula:

$$A = \frac{TruePositive + TrueNegative}{Total\ Predictions} \quad (1)$$

2) Precision: Precision refers to the ratio of true positive outcomes to the total predicted positive outcomes (both true and false positives). It is calculated as follows:

$$P = \frac{True\ Positives}{True\ Positives + False\ Positives} \quad (2)$$

3) Recall: Recall, also known as sensitivity, measures the proportion of actual positive cases that were correctly identified by the model. The formula used is:

$$R = \frac{True\ Positives}{True\ Positives + False\ Negatives} \quad (3)$$

4) F1-Score: The F1-score combines precision and recall into a single metric, especially useful in cases of imbalanced datasets. It is determined by calculating the harmonic mean of precision and recall.

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \quad (4)$$

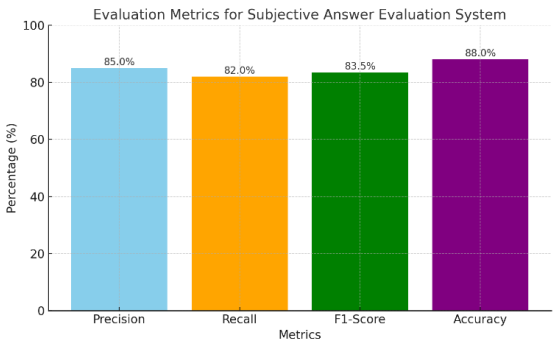

fig 2: Evaluation metrics for Subjective Answer Evaluation System.

B. *Cosine Similarity Performance for Subjective Answer Evaluation*

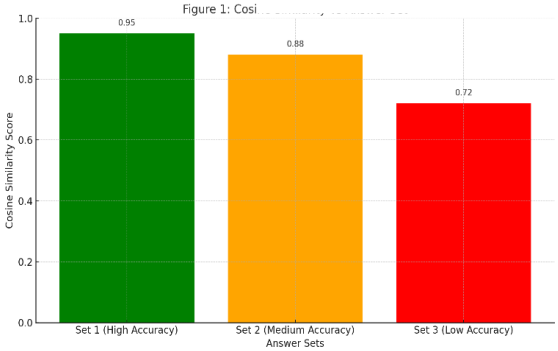

Fig 3: *Cosine Similarity Performance for Subjective Answer Evaluation*

Figure 3: Cosine Similarity vs Answer Set presents the performance of the system in evaluating subjective answers using the Cosine Similarity technique. This metric measures the angle between vectors representing the user's answer and the ideal answer, with values closer to 1 indicating high similarity and semantic alignment between the texts.

Observations

1. High Accuracy Set (Set 1):
   - Cosine Similarity Score: 0.95
   - Description: This set contains answers that are highly similar to the ideal answers, demonstrating excellent performance of the system for accurately written responses.
2. Medium Accuracy Set (Set 2):
   - Cosine Similarity Score: 0.88
   - Description: Answers in this set moderately align with the ideal answers, showing acceptable system performance for partially correct or less detailed responses.
3. Low Accuracy Set (Set 3):
   - Cosine Similarity Score: 0.72
   - Description: This set represents answers with minimal similarity to the ideal answers, highlighting the system's ability to identify less accurate or irrelevant responses.

V. CONCLUSION

The proposed system for Automatic Subjective Answer Evaluation effectively utilizes Natural Language Processing (NLP) techniques to analyze and evaluate textual responses provided by students. By leveraging Django as the web application framework and implementing key NLP processes such as tokenization, stop word removal, lemmatization, and similarity measurement using cosine similarity and semantic similarity, the system provides a robust, scalable, and automated solution for assessing subjective answers. The project demonstrates that the combination of rule-based methods and machine learning algorithms, including Logistic Regression and Support Vector Machines (SVM), enhances the classification accuracy of student responses. This approach reduces the dependency on manual grading, thereby improving efficiency, consistency, and fairness in evaluation.

The system achieves high accuracy for answers that closely resemble ideal responses, as evidenced by a cosine similarity score of 0.95 for highly accurate answers. By supporting a range of answers with varying levels of correctness, the system accommodates partial and alternative responses, making it adaptable to real-world educational environments. of educators, allowing them to focus more on personalized feedback and instructional improvements. This project contributes to the growing field of educational technology by demonstrating the practical application of AI and NLP in subjective answer evaluation. The findings and developed system highlight its potential to revolutionize traditional grading methods by making assessments more efficient, accurate, and equitable.

VI. FUTURE WORK

For further improvement, the following enhancements are suggested:

- Expanding the system to include adaptive feedback mechanisms that provide detailed suggestions for incorrect answers.
- Implementing support for multiple languages to broaden the system's usability across diverse educational contexts.

REFERENCES

[1] An Analysis Of Answer Evaluation System (IEEE Member)(2020) *1Birpal Singh J. Kapoor, 1Shubham M. Nagpure, 1Sushil S. Kolhatkar, 1Prajwal G. Chanore, 1Mohan M. Vishwakarma, 2Prof. Rohan B. Kokate*

[2] Subjective answers evaluation using machine learning and natural language processing (2023) *prof.mahalakshmi c v1, arymann sinha2, divin r3, hritesh r4, hrithik v4*

[3] Subjective answer assement using machine learning *August(2023) Chaithra L*1, Prof. Yogesha T*2*

[4] Automatic Evaluation of Descriptive Answers Using NLP and Machine Learning.(March 2022) *Prof. Sumedha P Raut1, Siddhesh D Chaudhari2, Varun B Waghole3, Pruthviraj U Jadhav4, Abhishek B Saste5*

[5] *Subjective Answers Evaluation Using Machine Learning and NLP MUHAMMADFARRUKHBASHIR1, HA*MZA ARSHAD1, ABDUL REHMAN JAVED 2,(Member, IEEE), NATALIA KRYVINSKA 3, AND SHAHAB S. BAND 4,(Senior Member, IEEE)(Deccember 2021