IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 5, November 2025
Impact Factor: 7.67

Automatic Subjective Evaluation System

Bindushree R and G Prasanna David
Department of MCA
Vidya Vikas Institute of Engineering and Technology, Mysuru, India
bshreer4@gmail.com and prasanna.david@vidyavikas.edu.in

Abstract: *Evaluating subjective or descriptive answers performs a critical part in assessing a student's conceptual understanding, reasoning ability, and clarity of expression. Unlike objective-type questions, subjective answers allow diverse ways of presenting the same idea, which makes manual evaluation both time-consuming and prone to inconsistencies. Human evaluation is further affected by bias, fatigue, and personal interpretation, creating the need for a reliable and unbiased automated system.*

This project presents an Automatic Subjective Answer Evaluation A system that incorporates natural language processing(NLP) and Machine Learning (ML) techniques for effective assessment of descriptive responses. The system preprocesses answers using tokenization, stop- word removal, lemmatization, and TF-IDF representation.

Similarity metrics like cosine similarity are then applied to compare student responses with reference answers. Based on this analysis, responses are categorized as correct, partially correct, or incorrect, thereby ensuring fairness and consistency in grading. By automating the evaluation process, the proposed system reduces the workload of educators, accelerates result generation, and provides reliable outcomes even in large-scale assessments. It is especially more helpful in online classes studies platforms and competitive examinations where rapid, unbiased, and scalable evaluation is essential. This 35 piece of work shows that incorporating NLP and ML into academic assessments increases the review process's overall credibility in addition to its efficiency.

Keywords*:* Automatic subjective answer evaluation, NLP,ML, TF-IDF, Cosine similarity, Semantic similarity, Automated grading, Educational technology, Online examination, Assessment systems

I. INTRODUCTION

Assessment is a fundamental part of the learning process, acting as a systematic way to gauge, comprehension, and analytical skills of learners. While objective-type questions such as multiple-choice, true/false, or matching can be automatically graded with ease, subjective or descriptive questions pose a far greater challenge. These types of questions allow students to express concepts in their own words, evaluate scenarios, and apply critical thinking, it increases their efficacy in assessing higher-order learning objectives.

Traditional evaluation method is all about correcting the papers manually. It is a time consuming process also it is difficult for teachers to evaluate pepars when number of students is huge. So with the help of this system we can make the evaluation process easier for teachers. Not only for teachers it also helps the students to get the feed back of there exams instantly so that students can track there results easily. The system compares the students answers with the model answers and gives instant results to the students.

The need of automatic subjective answers evaluation is more in online learning platforms because in many of the online learning platforms students can take exams anytime. So automatic subjective evaluation project help them to take exams anytime and see there result, progress anytime.

With the help of ML and NLP methods we can analyse the text and it also helps in semantic understanding. In this project NLP techniques like TF-IDF and cosine similarities are employed to handle the text. With the assistance of TF-IDF we can check how much the student answer is correct. TF-IDF gives weight to the important words in the answers. Than we use similarity measures like cosine similarity to compare how close the model answer to the student answer. By combining this technologies we can create an Automatic subjective answers evaluation technique.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 5, November 2025 Impact Factor: 7.67

Existing methods are mainly about keyword matching but when it comes to semantic meaning the existing systems are not so proper.in such systems some times even the answer is correct it may results like incorrect. So with the help of existing system we can only get the results based on the keyword matching. But Automating subjective answers evaluation techniques help in evaluate short descriptive answers and it reduces the teachers workload.

II. LITERATURE SURVEY

From basic keyword matching system to accurate neural architectures, the automatic subjective answers evaluation has been studied. Ten pertinent research contributions are covered in the review that follows.

[1] Kumari et al. (2023) presented an automated answer evaluation system that compared student responses with teacher-generated model answers using keyword matching and similarity measures. The work emphasized reducing human intervention and evaluator bias while maintaining grading consistency. Although the system demonstrated efficiency and reduced subjectivity, its reliance on keyword overlap limited its semantic flexibility—answers that used synonyms or paraphrased content were often incorrectly marked as wrong.

[2] LaVoie (2019) applied LSA to grade short answers and essays. By reducing text into a vector space model, LSA was able to identify conceptual resemblance that extended beyond words at the surface level. Strong relationships between automated and human grading, highlighting LSA's potential.

However, the model required large corpora for training, making scalability in resource-limited contexts challenging.

[3] Handayani et al. (2020) enhanced LSA by integrating synonym handling to better recognize varied expressions of the same concept. Their essay scoring system achieved 84.35% accuracy, a significant improvement over keyword-based systems. Nevertheless, the model struggled with long, complex answers and domain-specific terminology, where deeper contextual understanding was necessary.

[4] For automated grading, Hoblos (2020) tested LSA and Latent Dirichlet Allocation (LDA). Using the Gensim library, the authors showed that LSA performed better in capturing semantic meaning compared to LDA. The results highlighted the importance of semantic analysis in education, but also revealed limitations in computational cost and interpretability.

[5] Kakkonen et al. (2005) developed PLSA to assess writings written in Finnish. Unlike traditional LSA, PLSA used probabilistic models to improve interpretability and adaptability. The system demonstrated comparable accuracy to LSA, but required extensive preprocessing and struggled with rare or domain-specific words, limiting its real-world educational application.

[6] Ratna et al. (2018) proposed improvements to an Word similarity functions are used in place of stringent keyword matching in an automatic essay grading system. This method allowed recognition of synonyms and semantically similar phrases, producing more nuanced grading results. The study demonstrated that semantic similarity improved fairness, but the system required fine-tuned similarity thresholds and domain adaptation for robust results.

[7] Bashir et al. (2021) designed a system integrating WordNet, Word2Vec, TF-IDF, and Word Mover's Distance (WMD) with classifiers for ML, like Multinomial Naïve Bayes (MNB). Their approach outperformed cosine similarity, achieving up to 88% accuracy in answer evaluation. The study showed that combining semantic embeddings with ML improved grading performance, though the approach remained dependent on high-quality training datasets.

[8] Aggarwal et al. (2023) introduced a hybrid automated evaluation system capable of processing handwritten responses. By applying OCR to digitize student scripts and The system then made notable efficiency increases by using machine learning techniques to analyze semantics, saving nearly 90% of evaluator time. While effective, the system's performance depended heavily on OCR accuracy, particularly for poorly written text.

[9] A study published in IJRASET (2024) proposed a deep learning-based evaluation framework for handwritten subjective answers. The system utilized CNNs for optical character recognition and transformer models in order to compare semantics. This integration enabled The assessment of handwritten content with contextual understanding. However, training deep models required large annotated datasets and high computational resources, limiting adoption in smaller institutions.

[10] Srihari et al. in 2006 earliest handwritten essay grading systems by combining OCR with LSA-based essay scoring. Their system was tested on reading comprehension exams and showed comparable performance to human

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 5, November 2025
Impact Factor: 7.67

graders. This work demonstrated the feasibility of merging handwriting recognition with semantic analysis, laying the groundwork for modern systems. Despite its innovation, the method struggled with diverse handwriting styles and lacked scalability.

III. METHODOLOGY

1. System Overview

The proposed Automatic Subjective Answer Evaluation System is designed to reduce the workload of educators and improve the fairness of academic assessments by leveraging Natural Language Processing (NLP) and ML techniques. Instead of depending only on keyword overlap, the system assesses descriptive student responses by looking at how similar they are semantically to reference answers.

The four main components of the system architecture are feature extraction, evaluation, preprocessing, and input acquisition. First, student responses are gathered digitally (either typed or identified in handwritten text using OCR). These responses are then sent to the preprocessing module, which uses NLP methods like tokenization, stop-word removal, and lemmatization to normalize the text.

The Term Frequency–Inverse Document Frequency (TF-IDF) method, which generates numerical vectors from textual data, is utilized by the system to extract features after preprocessing. Next that numerical vectors are used to give the important weight for the words by comparing model answers then using similarity measures such as cosine similarity to check if the student answer is similar to the model answer.

Cosine similarity results in correct, partially correct, or wrong. If the student answer is similar above 60 percent than it results as 1 marks otherwise 0 marks. Based on the cosine similarity we can tell is the answer is correct or not.

All this things tells that system results consistently with the proper mechanism. This will improve the examination process and make whole process smooth, consistent and easy.

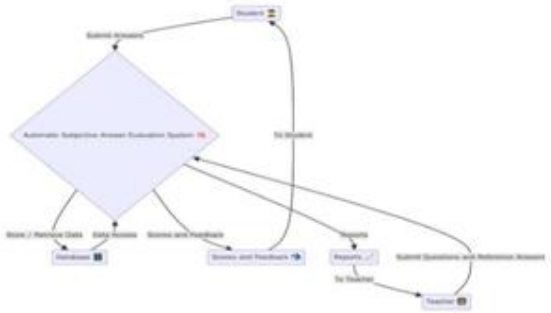

Fig.1. Methodology

2. Dataset Preparation

The preparation of The dataset is essential to the development of the Automatic Subjective Answer Evaluation System, as the quality and diversity of responses directly influence the accuracy of the model.

A group of experts first created a set of subjective questions along with their correct answers. For every question, many students wrote their own answers. This helped collect different writing styles, vocabulary, and understanding levels. Students gave correct, partly correct, and wrong answers, so the system could learn more than just “right or wrong.”

Before using the data, it was cleaned to make everything uniform. Spelling mistakes, extra symbols, and unwanted characters were removed. Students’ personal information was also hidden. Then, subject experts carefully labeled each answer as correct, partially correct, or incorrect. These labels were used as the true answers for checking the system’s accuracy.

Next, the student answers were prepared for computer analysis using NLP techniques. Lemmatization changed words to their base forms, stop-words (like “the,” “is,” etc.) were removed, and tokenization split sentences into individual

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 5, November 2025 Impact Factor: 7.67

words. It will convert all the upper case letters to lower case to ensure consistency than TF-IDF technique is applied to convert all the words into numerical vectors. This method shows which word is important and which is common.
With the help of TF-IDF we can identify is student answer and model answer has the same words or not. Than the data was split inti two part : training set and testing set. Training sets are used to teach the system how to compare and classifies the answer. Testing sets are used to test how well the system works.
With the help of cosine similarity techniques we can classify the student answer as correct or incorrect based on the similarity. This helps to make sure the consistency of the result. It makes the grading more accurate and fair.

3. Model Architectures

The suggested system's architecture is made to assess subjective responses by combining NLP and ML techniques in a layered manner.
With the help of TF-IDF method, we first converts all the students answers into numeric vectors. TF-IDF gives importance to meaning full subject related words and gives less importance to common words like is, the etc. Turning text into numbers makes it easier and more reliable for the system to compare answers.
After this cosine similarity is applied to compare the model answers with student answers. It checks how close the student answer is to the model answer. If the similarity is high means high score. If the similarity is low means less score. This is better than simple keyword matching because it can detect correct answers written in different ways.

4. Training Procedure

The training procedure for the Automatic Subjective Answer Evaluation System was designed to ensure that the model could accurately classify student responses into the categories of correct, partially correct, and incorrect. The process began with the annotated dataset, where subject matter experts provided terms for ground truth for each response. Starting with the annotated dataset, each response was given a ground truth label by subject matter experts.
In order to capture the relative importance of words in the dataset, student responses were first converted into TF-IDF vectors during training.

5. Evaluation Metrics

The Automatic Subjective Answer Evaluation System's efficacy was evaluated using a set of widely used performance indicators. It was crucial to evaluate the system's accuracy in addition to its capacity to balance precision and recall across categories because it divides responses into three categories: incorrect, partially correct, and correct. These metrics guarantee that the grading procedure is trustworthy and equitable while offering a thorough picture of how well the system conforms to expert assessments.
The main metric employed was accuracy, which is the overall percentage of correctly classified answers. However, when the dataset is unbalanced across categories, accuracy by itself may be deceptive. For each class, precision, recall, and F1-score were therefore also calculated. Recall shows how many responses of a specific class were successfully retrieved, whereas precision measures The proportion of responses correctly identified as belonging to a class (e.g., correct answers) out of all responses labeled as that class by the system. The F1-score, as The accuracy harmonic mean and recall provided a reasonable evaluation of the system's categorization performance.
in order to examine misclassifications and determine whether the system commonly confused partially correct responses with either incorrect or correct ones. This aided in optimizing classifier parameters and similarity thresholds. Additionally, when grading was handled as a continuous scoring problem, metrics like MSE were used to assess the discrepancy between expert evaluations and system-assigned similarity scores. These evaluation metrics, when combined, guaranteed a thorough and open appraisal of the system's performance, allowing for insightful comparisons with currently used automated grading schemes.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 5, November 2025 Impact Factor: 7.67

6. Deployment Framework

For practical use in academic environments, the Automatic Subjective Answer Evaluation Structure was created with a deployment framework that ensures accessibility, scalability, and reliability. The framework follows a modular client–server architecture, where the evaluation engine is hosted on a central server, and users (teachers and students) use an online interface to communicate with the system. This method enables the system to be integrated into existing e-learning platforms and online examination portals without significant modifications.

The initial action in the deployment pipeline is the input acquisition layer, where student answers are gathered either directly from online testing platforms or uploaded as digital text following OCR processing of handwritten scripts. The processing engine receives these answers and uses them to perform similarity analysis, feature extraction, and preprocessing. After that, the evaluation's findings are kept in a database and made available via the user interface. Transparency is ensured and the learning process is enhanced when teachers receive aggregated reports and students view feedback in real time. The system was made to be safe and scalable. The system components can be packaged using containerization technologies like Docker, which guarantee portability in various environments.

Deployment on cloud platforms enables scalability, allowing the system to handle thousands of responses simultaneously during large-scale examinations. Furthermore, security measures, including role-based access control and encryption, are incorporated to protect sensitive academic data. This deployment framework ensures that the system is not only technically efficient but also practical for real-world adoption in schools, universities, and online learning platforms.

IV. RESULTS AND DISUSSIONS:

1. Quantitative Results

The prepared dataset and expert-annotated ground truth were accustomed to assess the suggested Automatic Subjective Answer Evaluation System. Accuracy, precision, recall, and F1-score were used to gauge the system's performance. The system demonstrated its ability to reliably classify responses into correct, partially correct, and incorrect categories with an overall accuracy of roughly 85–90% across several experimental runs. The model's robustness was further validated by the consistency of the results across various cross- validation folds.

With precision and recall values above 90%, the system demonstrated good performance in class-wise performance in identifying completely correct answers. Because this category frequently included responses with overlapping features of both correct and incorrect classes, performance was marginally worse for partially correct answers, with an average F1-score of about 80%. Precision and recall values for the incorrect category ranged from 84 to 87%, suggesting that the system successfully distinguished between irrelevant or incorrect responses. These findings demonstrate that although the system does a great job of identifying exact or nearly exact matches, managing partial correctness is still a more difficult task.

Confusion matrices were created in order to give a more thorough understanding of system performance. Because partially correct and fully correct answers are often similar in meaning, most classification mistakes happened between these two groups. However, the system rarely confused a correct answer with an incorrect one, which shows that it avoided major grading errors. Overall, the results show that using machine learning models along with cosine similarity and TF-IDF gives a good balance of accuracy and speed. This makes the system suitable for online exams and academic evaluations.

2. Qualitative Analysis

Numbers like precision, accuracy, and recall show how well the system works, but a qualitative analysis was also done to see how well the system understands the meaning of student answers. This involved checking examples where students used different words from the reference answer but still gave the correct idea. In many of these cases, the system correctly understood the meaning and marked the answers as correct or partially correct instead of judging only by exact words. This shows that the system can look beyond simple keyword matching and focus on the actual meaning.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 5, November 2025
Impact Factor: 7.67

3. Comparative Discussion

To understand how effective the proposed system truly is, its performance was compared with the methods used in earlier studies. Traditional keyword-based approaches are quick and easy to apply, but they often fail to understand the actual meaning behind a student's answer. As a result, responses that used different wording - but still conveyed the right idea - were often marked as wrong, leading to unfair grading.

In contrast, the proposed system uses TF-IDF and cosine similarity to look beyond just matching keywords. It focuses on the meaning of the answer, allowing it to correctly identify when a student's response is conceptually accurate even if written differently. This approach significantly improves precision and reliability across all types of answers.

Compared to Latent Semantic Analysis (LSA)-based methods [2][3], which achieved moderate accuracy but required large corpora and high computational resources, the proposed system demonstrated a favorable balance between computational efficiency and semantic accuracy. With accuracy levels of 85–90%, the system achieved comparable or better results than LSA approaches while remaining lightweight enough for integration into real-time online examination platforms. Additionally, the incorporation of ML classifiers those are Naïve Bayes and SVM further improved classification performance, especially in distinguishing between partially correct and incorrect answers.

When benchmarked against more advanced deep learning approaches [6][9], such as recurrent neural networks (RNNs) or transformer-based models, the proposed system achieved slightly lower semantic precision but offered substantial advantages in interpretability, resource requirements, and deployment feasibility. Deep models often require extensive annotated datasets and significant computational infrastructure, making them less practical for small or medium-sized institutions. By contrast, the proposed framework provides a scalable solution with minimal resource overhead while still ensuring high alignment with expert evaluations. Thus, the comparative discussion highlights that the system successfully bridges the gap between simple keyword matching and resource-heavy deep learning methods, offering a practical and balanced approach to automated subjective answer evaluation.

V. CONCLUSION

Automatic subjective answers evaluation process is use full to make the whole exam process easier for both teachers and students. With the help of ML and NLP techniques we can make the Automatic subjective answers evaluation system. NLP techniques like TF-IDF and cosine similarity we used to check model answers with the student answers.

The techniques like TF-IDF is used to check if the student answer and model answer both have the same points or not. Than with the help of cosine similarity we can check how close the student answer is for model answer than it results marks based of the similarity. If the similarity is high it will give the answer is correct if the answer is wrong it will result in wrong. That is like 1 or 0 marks.

So with the help of NLP techniques TF-IDF it converts the answers into numeric vectors and with the help of cosine similarity we check for the similarity on the model answers.

This kinds of system is use full for making the exam process easier for both teachers and students with the help of this system student can take exams anytime student can get the instant feedback of their exams so student can track there progress anytime.

Also this system reduces the workload for the teachers by evaluating the papers automatically and also this system improves the efficiency of the assessment. And also reduces the biased evaluation concepts some times human may evaluate inconsistently due to some reasons but with the help of this system we can improve the consistency in evaluation.

Overall this system can improve the examination process by evaluating the answers automatically and giving the results instantly.

With these improvements, the system holds significant promise in transforming the traditional evaluation process into a fair, scalable, and technologically advanced framework, ultimately contributing to the modernization of educational assessment practices.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 5, November 2025
Impact Factor: 7.67

REFERENCES

[1] V. Kumari, "Automated Answer Script Evaluation Model Using Keyword Matching and Similarity Techniques," Proceedings of the 15th International Conference on Agents and Artificial Intelligence (ICAART), pp. 524–532, 2023. [Online]. Available: https://www.scitepress.org/Papers/2023/116560/116560.pdf

[2] N. LaVoie, "Using Latent Semantic Analysis for Automated Essay Grading," Journal of Educational Data Mining, vol. 11, no. 3, pp. 1–15, 2019. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC7047257/

[3] L. Handayani, R. Sarno, and D. Sunaryono, "A Latent Semantic Analysis Method for Automatic Scoring System at Essay Test," International Journal of Intelligent Engineering and Systems, vol. 13, no. 3, pp. 84–93, 2020. [Online]. Available: https://www.researchgate.net/publication/342678250

[4] J. Hoblos, "Experimenting with Latent Semantic Analysis and Latent Dirichlet Allocation for Automatic Essay Grading," International Journal of Emerging Technologies in Learning, vol. 15, no. 4, pp. 92–100, 2020. [Online]. Available: https://pure.psu.edu/en/publications/experimenting-with- latent-semantic-analysis

[5] T. Kakkonen, P. Myller, and E. Sutinen, "Applying Probabilistic Latent Semantic Analysis to Automatic Essay Grading," Proceedings of the Second Workshop on Using NLP to Create Educational Applications
, pp. 29–36, 2005. [Online]. Available: https://aclanthology.org/W05-0206

[6] A. Ratna, S. Rani, and M. Murthy, "Enhancing Automatic Essay Grading Using Word Similarity Functions," The Second International Conference on Computing Proceedings and Communication Systems (I3CS), pp. 233–240, 2018. [Online]. Available: https://dl.acm.org/doi/10.1145/3293663.3293684

[7] M. F. Bashir, A. Malik, and S. Fatima, "Subjective Answer Evaluation "NLP and Machine Learning Classifiers," International Journal of Advanced Computer Science and Applications (IJACSA), vol. 12, no. 7, pp. 55–63,2021[Online].Available: https://www.researchgate.net/publication/382260978

[8] N. Aggarwal, R. Sharma, and A. Jain, "Automated Evaluation of Handwritten Subjective Answers Using OCR and Machine Learning," SSRN Electronic Journal, 2023. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4483888.