Contour Launchpad – Summer Internship Program 2025

Project Report

SMARTGRADE

SUBJECTIVE ANSWER EVALUATION SYSTEM

Qurat-ul-Ain Akhter

Employee ID: 132347

1. Executive Summary

This report presents *SMARTGRADE*, an AI-powered **Subjective Answer Evaluation System** developed to address the growing challenges of grading open-ended academic responses in a fair, consistent, and scalable manner. Traditional manual grading methods suffer from bias, inconsistency, fatigue, and are highly resource-intensive—especially in large-scale or resource-constrained educational environments. With the rising demand for digital learning and online assessments, the need for an automated evaluation system is more critical than ever.

SMARTGRADE offers a robust technical solution that combines **semantic similarity**, **keyword relevance**, and **grammatical correctness** to evaluate short to medium-length subjective answers. It leverages **transformer-based NLP models** (like BERT and RoBERTa), machine learning techniques (Gradient Boosting, SVR, etc.), and rule-based logic to predict accurate and explainable scores. The system provides two key evaluation modes—**single-response grading** and **bulk evaluation** via CSV upload—through an interactive frontend developed with React.js and a Flask-based backend.

Key differentiators of SMARTGRADE include:

- Multi-dimensional scoring with component-wise breakdown.
- Support for **handwritten answer evaluation** via OCR.
- Integration readiness with Learning Management Systems (LMS).
- Customizable scoring and feedback generation.
- Future-ready features like **plagiarism detection** and **multilingual support**.

Market analysis highlights a strong and growing demand for such systems, especially in developing regions, higher education institutions, and online course platforms. Unlike partial competitors such as Gradescope or EvalAI, SMARTGRADE provides a **holistic** and **explainable** evaluation framework that makes it not only a grading tool but also a **learning enhancement platform**.

The system was developed using synthetically generated student data from Kaggle-sourced questions. Gradient Boosting emerged as the most accurate ML model with a Mean Absolute Error (MAE) of 2.86, outperforming traditional rule-based methods. The Agile methodology ensured iterative development, continuous validation, and efficient resource use.

Challenges included the lack of real-world data and limited automation in data generation, which impacted training accuracy. However, these are addressed in the **future roadmap**, which includes collecting real student data, implementing plagiarism detection, automating feedback, and conducting pilot deployments.

In conclusion, SMARTGRADE represents a significant step toward transforming educational assessment by combining AI's power with the pressing need for scalable, fair, and insightful grading systems. Its adaptability, accuracy, and student-centric feedback mechanism position it as a strong contender in the EdTech landscape.

2. Table of Contents

3. Problem Statement and Objectives

Evaluating subjective and open-ended answers in academic settings continues to be a complex and time-consuming challenge. Unlike objective questions, subjective answers require deeper understanding and interpretation, making their assessment prone to human error, inconsistency, and bias. Teachers and examiners, especially in large-scale settings, struggle to evaluate hundreds or thousands of responses in a fair and timely manner. This bottleneck often leads to delayed results, increased workload, cognitive fatigue, and compromised grading quality.

The situation is further exacerbated in regions where resources are limited and standardized rubrics are lacking. Many educational institutions still rely on manual grading systems where personal perceptions, fatigue, or unintentional favoritism can significantly impact the evaluation outcome. Such inconsistencies can damage students' confidence and hinder their learning progress. With the rise of online education platforms, remote learning, and e-assessments, the need for scalable and reliable automated grading systems has become more urgent than ever.

The relevance and criticality of this problem have been widely recognized in academic and industry research. A study published by SCITEPRESS emphasized that automatic evaluation of subjective answers using NLP can significantly reduce workload and increase grading fairness by modeling semantic understanding rather than surface-level keyword matching (Abhay et al., 2023). Similarly, a paper published in IEEE explored the potential of machine learning and semantic similarity to predict grades for open-ended responses with high accuracy, confirming the usefulness of transformer-based embeddings in educational assessment (Manna & Das, 2021). Research published in ScienceDirect further confirmed that integrating grammatical assessment with semantic models improves not only grading precision but also the quality of feedback to students (Li et al., 2023). Moreover, the IRJMETS study emphasized the practicality of such automated systems in developing regions, where access to trained evaluators and scalable solutions is often limited (Shaikh et al., 2024).

These studies collectively validate the existence of a significant problem in current educational practices and support the development of automated, intelligent grading systems.

The objective of this project is to design and develop an AI-powered system titled SMARTGRADE: Subjective Answer Evaluation System that aims to resolve these challenges. The specific, measurable goals of the project are as follows:

- To create a scalable, automated evaluation system that can analyze and score short to medium-length subjective answers based on semantic similarity, keyword coverage, and grammatical correctness.
- To develop and integrate a multi-metric scoring mechanism using pre-trained NLP models and statistical techniques to generate scores aligned with human evaluation.
- To train and test machine learning models such as Gradient Boosting, Linear Regression, and SVR to identify the most accurate prediction method.
- To provide both a single-response evaluation tool and a bulk-response evaluation interface that allows educators to input reference answers and student responses.

- To ensure explainable scoring by breaking down the total score into individual components (semantic, keyword, and grammar scores) and visualizing feedback.
- To validate the proposed solution using synthetic student data generated with AI tools, simulating a real-world educational environment.

These objectives align directly with the central goal of improving grading efficiency, consistency, and scalability, particularly in education systems where manual evaluation continues to limit progress.

4. Market and Competitive Analysis

4.1. Market Trends

The demand for automated assessment tools in the education sector is growing rapidly due to the increasing shift towards digital learning platforms and remote education. Market trends indicate a strong inclination toward AI-driven evaluation systems that offer scalability, consistency, and efficiency. The global EdTech market is projected to reach USD 404 billion by 2025, with intelligent assessment solutions forming a significant segment of this growth. Institutions and examination boards are increasingly adopting AI-based systems for evaluating assignments, especially in higher education and skill-based assessments.

4.2. Market Demand

There is a clear market demand for solutions like SMARTGRADE. The growing number of students, the popularity of MOOCs (Massive Open Online Courses), and digital university programs have made it necessary to process large volumes of open-ended responses without delays. Instructors and exam administrators need tools that not only score answers accurately but also provide explainable feedback that supports learning. This demand is further accelerated in developing regions where the teacher-to-student ratio is critically low and manual grading becomes a bottleneck in education delivery.
SMARTGRADE is also built to handle evaluations from handwritten exam papers. Using Optical Character Recognition (OCR) technology, answer sheets written by hand can be scanned and processed into machine-readable formats, thus broadening the applicability of the system to conventional pen-and-paper examinations. This feature enhances its value in schools and institutions that still rely on offline assessments but seek to modernize their grading practices.

4.3. Competitors

Several competitors currently operate in the space of automated answer evaluation. Tools like Gradescope, EvalAI, and OpenAI's GPT-based evaluators offer partial automation for short-answer or code-based grading. However, these tools often focus on limited domains or require significant customization for subjective answer evaluation in theoretical subjects like Data Science. Gradescope, for instance, allows semi-automated grading but still depends heavily on predefined rubrics. OpenAI-based solutions can evaluate free-form text but often lack integration of grammar and keyword-level analysis.

SMARTGRADE differentiates itself by offering a holistic scoring framework that combines semantic similarity, keyword relevance, and grammatical correctness. Furthermore, it provides transparency through component-wise score breakdowns, which are not commonly available in competing tools. The use of AI-generated student responses during training also gives SMARTGRADE an edge in terms of data diversity and adaptability. Moreover, its ability to handle scanned handwritten inputs expands its reach to offline classrooms and rural education systems, making it more inclusive than most alternatives.

4.4. Stakeholder Mapping

In terms of stakeholder mapping, the primary stakeholders include educational institutions (schools, colleges, and universities), teachers and examiners, EdTech companies, and governmental bodies involved in educational assessment. Secondary stakeholders include students who benefit from fairer, more timely evaluations, and researchers in the field of AI in education.

4.5. Target Market

The target market for SMARTGRADE includes:

- Universities and colleges conducting large-scale theory-based exams
- Online course platforms offering data science and computer science content
- Education boards in developing countries looking to scale grading operations
- EdTech startups seeking AI-powered back-end evaluation services
- Traditional schools conducting handwritten exams but willing to digitize grading via scanning and OCR

This analysis confirms that there is strong market potential for a system like SMARTGRADE, particularly because it addresses both quality and scalability concerns in subjective evaluation.

5. Proposed Solution

5.1. Proposed Technical Solution

To address the challenges associated with the manual evaluation of subjective and open-ended responses, a comprehensive Automated Answer Evaluation System is proposed. This system leverages advanced Natural Language Processing (NLP) and Machine Learning (ML) techniques to evaluate textual responses in a scalable, consistent, and unbiased manner. The solution is designed to function independently or integrate with existing Learning Management Systems (LMS), educational platforms, or examination software.

5.2. Product/Offering Description

The Automated Answer Evaluation System comprises four key modules to ensure accurate, scalable, and unbiased assessment of subjective responses:

1. Multi-Dimensional Evaluation Framework
   Evaluates answers across various dimensions including content relevance through semantic similarity, conceptual understanding using deep language models, language quality (grammar, spelling, punctuation), coherence and structure, and keyword matching for key term identification.

2. AI and NLP Integration

   Leverages advanced NLP techniques such as transformer models (e.g., BERT, RoBERTa) for deep contextual understanding, Semantic Textual Similarity (STS) to compare answers, Named Entity Recognition (NER) to check concept inclusion, and text classification for scoring and categorization.

3. Automated Scoring and Feedback Generation

   Enables dynamic scoring based on customizable rubrics, automated feedback highlighting strengths and weaknesses, and explainable evaluation with justification for each score.

4. Plagiarism and Originality Detection

   Detects copied content by comparing answers against existing databases, academic sources, and the web to ensure originality.

5.3. Unique Selling Point (USP)

The unique strength of SMARTGRADE lies in its multi-dimensional and explainable evaluation framework, which goes beyond simple keyword matching. It combines semantic similarity, keyword relevance, and grammatical accuracy to provide a deeper, more accurate assessment of student responses. Unlike many systems limited to typed answers, SMARTGRADE supports handwritten responses via OCR and delivers clear, interpretable score breakdowns. Its scalability, LMS integration capability, and customizable scoring logic make it a flexible and market-ready solution.
Moreover, SMARTGRADE includes features that most competitors overlook—automated feedback generation and plagiarism detection. These additions transform the system into not just a grading tool, but a learning support platform. By offering constructive feedback and ensuring content originality within the same system, SMARTGRADE empowers both educators and students while maintaining academic integrity.

5.4. Scope of the Project

The primary focus was on building a core prototype of the Subjective Answer Evaluation System using AI-generated data due to the unavailability of real-world student responses. Data collection involved scraping over 350 Data Science questions and answers from Kaggle, followed by generating 10 diverse student-like responses per question using AI tools. The system evaluates these responses across three key dimensions—semantic similarity (using transformer-based embeddings and cosine similarity), keyword overlap (via Jaccard similarity), and grammatical quality (via syntactic checks using spaCy). Two scoring approaches were implemented: one based on weighted rule-based aggregation and the other using machine learning models, where Gradient Boosting outperformed others. A basic frontend with both single and bulk evaluation features was also developed to demonstrate system usability.

6. Implementation Strategies

The development of the Subjective Answer Evaluation System was guided by an Agile methodology, enabling flexible and incremental progress through iterative phases of data acquisition, processing, model experimentation, and full-stack integration.

The first phase focused on data collection, where two publicly available Kaggle datasets containing 158 and 199 Data Science questions and their respective answers were selected. Due to the impracticality of acquiring real student responses within the project duration, AI tools were employed to generate 10 simulated student-like answers for each question. These responses varied in quality, covering correct, partially correct, and incorrect examples. Alongside the responses, the AI was instructed to assign corresponding scores on a scale from 0 to 1. The data was stored in two separate datasets—one containing the answers and the other the scores—which were later merged into a unified dataset containing questions, reference answers, generated responses, and their evaluation scores.

The second phase involved implementing the core evaluation metrics. For semantic similarity, the pre-trained sentence transformer model "all-MiniLM-L6-v2" was used to generate vector embeddings of both the reference answer and the student response. Cosine similarity between these vectors produced the semantic score. To assess keyword similarity, preprocessing techniques such as lowercasing, punctuation removal, normalization, tokenization, stopword elimination, and lemmatization were applied. The cleaned texts were then evaluated using Jaccard similarity, giving a measure of lexical overlap. Grammar quality was scored using spaCy's language parser, which detected grammatical issues such as missing subjects or verbs, incorrect punctuation, and passive voice. These issues were quantified and normalized into a score ranging from 0 to 1 using a custom Python function.

For score prediction, two approaches were explored. The first was a rule-based method, which combined the three scores (semantic, keyword, grammar) using a weighted average: 70% semantic, 25% keyword, and 5% grammar. The second method involved training five machine learning models—Linear Regression, Random Forest, Gradient Boosting, SVR, and XGBoost—using the three scores as input features to predict the actual score. Among these, Gradient Boosting yielded the best results with a Mean Absolute Error (MAE) of 2.86, outperforming the rule-based approach, which had an MAE of 3.43. The Gradient Boosting model was thus selected for final deployment.

In the final phase, a user interface was developed using React.js for the frontend and Flask for the backend. The backend handled model inference and evaluation logic, while the frontend offered a clean and interactive UI for end-users. These components were integrated via RESTful APIs, enabling smooth communication between the client and server.

The web application provides two primary functionalities:

The single-answer evaluation module allows users to input a question, reference answer, student response, and maximum marks, and receive a detailed breakdown including semantic score, keyword score, grammatical score, and the final predicted score.

The bulk evaluation feature accepts a CSV file containing multiple student responses and generates a comprehensive result sheet including roll numbers, all sub-scores, and final predicted scores, with an option to download the output as a CSV.

Technologies used in the project include Python as the core programming language, along with libraries such as pandas, numpy, scikit-learn, spaCy, sentence-transformers, and matplotlib for development and evaluation. React.js was used for building the user interface, and Flask was used to build the backend API. The system was developed and deployed using a personal computer with an

entirely open-source software stack. The Agile workflow ensured continuous testing and validation, effective resource utilization, and timely project completion.

7. Risks and Challenges

The implementation of the Subjective Answer Evaluation System encountered several challenges, the foremost being the unavailability of real-world student response data. Due to the limited timeframe, it was not possible to collect authentic responses from students across different proficiency levels. To address this, AI tools were used to generate student-like responses along with corresponding scores. However, these tools lacked bulk processing capabilities, requiring each question to be manually inputted and its responses and scores to be extracted individually. This not only made the data collection process tedious and time-consuming but also introduced risks of data inconsistency and increased the burden of manual verification when storing responses and scores in separate CSV files.

Another significant challenge was achieving high prediction accuracy. Since the training data was synthetically generated and the models used were pre-trained general-purpose transformers, the semantic understanding and scoring were not always as precise as desired. Despite fine-tuning and preprocessing efforts, the system's performance, though acceptable for a prototype, showed a Mean Absolute Error (MAE) of 2.86 using the best-performing Gradient Boosting model—indicating room for improvement. These limitations highlight the need for better training data and more domain-specific models to enhance accuracy and generalizability in future iterations of the system.

8. Future Plans/Recommendations

Following the initial development of SMARTGRADE, several steps are planned to enhance its effectiveness, scalability, and real-world integration. A primary focus is improving the model's scoring accuracy by incorporating real student response data. Collaborating with educational institutions to gather authentic and diverse datasets will help reduce the reliance on AI-generated responses and improve prediction reliability.

Additionally, SMARTGRADE will be enhanced with critical functionalities such as automated feedback generation, which will provide students with constructive and actionable insights, and plagiarism detection, ensuring academic integrity by checking originality against web sources, databases, and peer submissions. Upgrading existing models with subject-specific fine-tuning will also be prioritized to boost contextual understanding.

Next 1 Month:

- Begin integration of automated feedback generation.
- Improve semantic similarity preprocessing and normalization.

Next 2–6 Months:

- Implement plagiarism detection module.
- Initiate data collection from real students in collaboration with educators.
- Conduct internal accuracy testing and calibration of evaluation metrics.

Next 7–12 Months:

- Fine-tune transformer models on collected educational data.
- Enable multilingual support for answer evaluation.
- Launch pilot deployments within small classroom or institutional settings.
- Prepare system for LMS integration and institutional use.

By executing these plans in phases, SMARTGRADE aims to transition from a working prototype to a scalable, accurate, and widely adoptable educational assessment solution.

9. Appendices

9.1. Dataset Snapshot

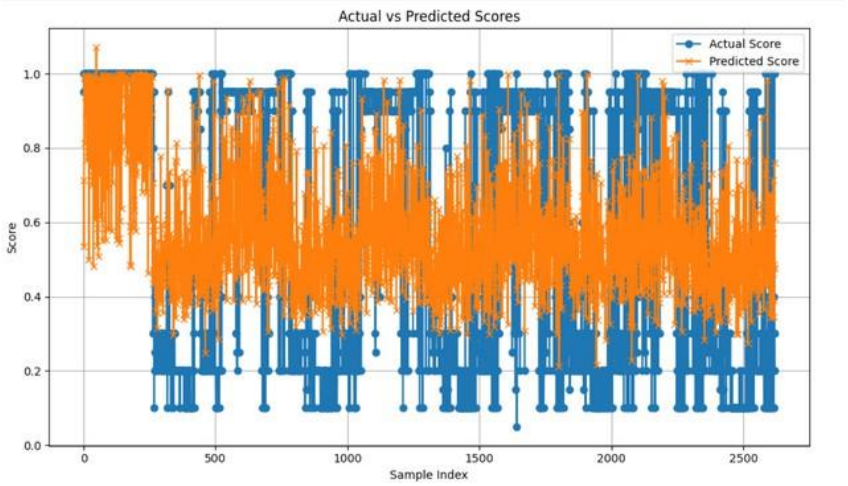
	Question	Answer	Response	semantic_similarity	keyword_score	grammar_score	Score
0	What is data science?	Data science is an interdisciplinary field tha...	Data science is a field that combines statisti...	0.866848	0.176471	0.5	1.00
1	What are the key steps in the data science pro...	The key steps typically include problem defini...	The key steps in the data science process incl...	0.651658	0.121951	1.0	0.95
2	What is the difference between supervised and ...	Supervised learning involves training a model ...	Supervised learning uses labeled data to train...	0.915449	0.342105	1.0	1.00
3	Explain the bias-variance tradeoff.	The bias-variance tradeoff is the balance betw...	The bias-variance tradeoff is a concept in mac...	0.915531	0.346154	1.0	1.00
4	What is feature engineering?	Feature engineering is the process of selectin...	Feature engineering is the process of creating...	0.875637	0.250000	1.0	1.00
...	...	...	...	...	...	...	...
2615	What are the common density estimation techniq...	Common techniques include histogram-based meth...	Decision trees are commonly used density estim...	0.597383	0.157895	0.2	0.10
2616	What is the Gaussian Mixture Model (GMM)?	Gaussian Mixture Model (GMM) is a probabilisti...	GMM is faster than K-means because it uses few...	0.401086	0.050000	1.0	0.40
2617	What is the Expectation-Maximization (EM) algo...	The Expectation-Maximization (EM) algorithm is...	EM is especially useful when direct optimizati...	0.627378	0.142857	1.0	1.00
2618	What is the difference	Generative models learn the	The main difference is that	0.639115	0.263158	1.0	0.30

9.2. Model Performance Summary Table

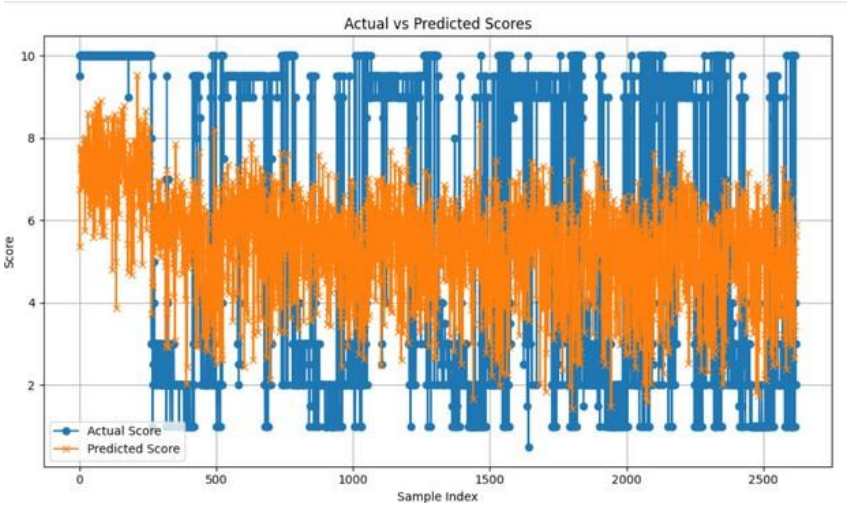
Model	Mean Squared Error (MSE)	R^2 Score
Linear Regression	0.1204	0.1309
Random Forest	0.1477	-0.0665
Gradient Boosting	**0.1164**	**0.1599**
Support Vector Regressor (SVR)	0.1280	0.0761
XGBoost	0.1567	-0.1313

9.3. MAE Comparison Line Plot

By Gradient Booting Model:



By Weighted Average approach:



10. References

Abhay, S., Sharma, D., & Chauhan, D. S. (2023). Automatic evaluation of subjective answers using NLP and similarity models. In Proceedings of the 15th International Conference on Agents and Artificial Intelligence.
SCITEPRESS. http s://www.scitepress.org/Papers/2023/116560/116560.pdf

Shaikh, S., Patel, A., & Joshi, P. (2024). AI Based Subjective Answer Evaluation System. International Research Journal of Modernization in Engineering Technology and Science (IRJMETS).
http s://www.irjmets.com/uploadedfiles/paper/issue_8_august_2024/60829/final/fin_irjmets1723039704.pdf

Manna, A., & Das, P. (2021). Evaluation of Subjective Answers Using Semantic Similarity and Supervised Learning. 2021 International Conference on Computer Communication and Informatics (ICCCI).
IEEE. http s://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9627669

Li, Z., Huang, J., & Li, Y. (2023). Integrating Grammar and Semantics for Better Feedback in Automated Text Scoring. Computers in Human Behavior Reports, Elsevier.
http s://doi.org/10.1016/j.chbr.2023.100230