Subjective Answers Evaluation

Prof. Chinmay Raje[1], Prasad. D. Jagdale[2], Prathamesh Pandhare[3], Wrushabhkumar Bangar [4]
Pritam Jadhav[5]
[1]*Professor SKN Sinhgad Institute of Technology and Science, Lonavala*
[2,3,4,5]*Student SKN Sinhgad Institute of Technology and Science, Lonavala*

***Abstract*—Subjective questions and responses provide an open-ended assessment of a student's understanding, allowing them to express their knowledge in a personalized and conceptual manner. However, the manual evaluation of such answers is often time-consuming, inconsistent, and prone to bias. This project proposes an automated system for evaluating subjective answers using Machine Learning (ML) and Natural Language Processing (NLP) techniques. The system utilizes various NLP methods and models such as Word2Vec, WordNet, Word Mover's Distance (WMD), Cosine Similarity, Term Frequency-Inverse Document Frequency (TF-IDF), and Multinomial Naive Bayes (MNB) to analyze and score answers. By comparing student responses to reference answers on the basis of semantic similarity and keyword relevance, the model predicts a score with high accuracy. The system aims to improve grading consistency, reduce evaluation time, and enhance the overall efficiency of academic assessments. Experimental results show that the WMD technique performs better than Cosine Similarity in maintaining semantic integrity, and with sufficient training, the machine learning model is capable of functioning autonomously.**

***Index Terms*—Subjective Answer Evaluation, Natural Language Processing (NLP), Machine Learning (ML), Word2Vec, Word Mover's Distance (WMD), Cosine Similarity, TF-IDF, Multinomial Naive Bayes (MNB), Semantic Similarity, Automated Grading, Educational Technology, Text Preprocessing.**

I. INTRODUCTION

Evaluating subjective papers by hand is a difficult and time-consuming undertaking. One of the biggest obstacles to employing artificial intelligence (AI) for the subjective paper analysis process is a lack of comprehension and acceptance of the findings. There have been several attempts to use computer science to grade students' responses. To do this, the majority of the job, however, makes use of conventional counts or certain terms. Additionally, vetted data sets are also lacking. In order to automatically evaluate descriptive responses, this paper suggests a novel method that makes use of a variety of machine learning, natural language processing, and toolkits, including Wordnet, Word2vec, word mover's distance (WMD), cosine similarity, multinomial naive bayes (MNB), and term frequency-inverse document frequency (TF-IDF). Answers are assessed using keywords and solution statements, as well as a machine learning algorithm.

II. LITERATURE SURVEY

The evaluation of subjective answers using automated systems has gained significant attention due to the challenges and limitations of manual grading. Several studies have explored various approaches combining machine learning (ML) and natural language processing (NLP) techniques to address this issue. Dr. Neha R. Sharma and Aakash Patil emphasized the use of semantic-based models such as BERT and Sentence Transformers, highlighting their superiority over traditional keyword-based methods [1]. Dr. Meenal Joshi and Harshad Deshmukh proposed a hybrid approach integrating Word2Vec with Word Mover's Distance (WMD) to measure semantic similarity between student and reference answers, achieving results closely aligned with manual scoring [2]. Prof. S. R. Kulkarni and Priya Wagh combined BERT embeddings with Multinomial Naive Bayes, demonstrating improved accuracy in grading tasks by capturing the context of answers more effectively [3]. Similarly, Dr. Lata Verma and Omkar Pawar introduced a dual-phase system using TF-IDF for feature extraction and deep learning for classification, showing enhanced performance and reduced bias [4]. Additionally, Muhammad A. Khan and Sanya Nair proposed a framework using BERT along with WordNet-based similarity metrics, which proved effective in recognizing conceptually correct yet differently phrased answers [5]. These studies collectively demonstrate that leveraging advanced NLP models and semantic similarity techniques significantly improves the fairness, consistency, and efficiency of subjective answer evaluation systems.

III. OBJECTIVES

The primary objectives of this project are as follows:
1. Automate the Evaluation Process: Build an intelligent system that evaluates subjective answers automatically, reducing the need for human intervention.
2. Improve Accuracy and Consistency: Traditional methods of grading subjective answers are prone to human bias and inconsistency. This project minimizes these issues by using machine learning models that analyze semantic meaning.
3. Contextual Understanding: By employing NLP techniques like Word2Vec and Word Mover's Distance, the system evaluates the context of words, ensuring fair evaluation even for paraphrased or creatively structured answers.
4. Efficiency in Evaluation: Automating the grading process will significantly reduce the time taken to evaluate large sets of subjective answers, making it scalable for educational institutions.
5. Adaptability: The system should be flexible enough to handle various subjects and domains, allowing its use in different academic disciplines.

IV. SYSTEM ARCHITECTURE

The proposed system architecture for the evaluation of subjective answers using Machine Learning and Natural Language Processing is designed to handle the entire process of answer analysis from input processing to score generation. The architecture is modular, ensuring each component performs a dedicated function efficiently and can be upgraded independently.

Architecture Overview

1. Input Module: This module accepts user input in the form of text-based answers or scanned handwritten responses (using OCR for digitization, if required).
2. Pre-processing Module: It cleans the input data by removing stopwords, performing lemmatization, and tokenizing the text to ensure consistency and uniformity in analysis.
3. Feature Extraction Module: This component uses NLP techniques such as TF-IDF, Word2Vec, or BERT to transform textual data into vector representations, enabling the system to understand semantic similarities.
4. Similarity Analysis Module: This module compares student answers with standard reference answers using similarity measures like Cosine Similarity, Word Mover's Distance (WMD), or Jaccard Index.
5. Classification and Scoring Module: A trained machine learning model (e.g., Multinomial Naive Bayes, SVM, or Neural Networks) predicts scores based on extracted features and similarity scores.
6. Feedback and Output Module: Finally, the system generates a numerical score along with optional feedback highlighting missing key points or grammatical corrections.

This architecture ensures a streamlined and scalable solution that accurately assesses subjective responses by leveraging the strengths of modern NLP and ML algorithms.

V. IMPLEMENTATION

The implementation of the Subjective Answer Evaluation System is centered around the integration of Natural Language Processing (NLP) and Machine Learning (ML) modules, designed to automatically assess descriptive answers. The system is built using Python and leverages key libraries such as NLTK, Word2Vec, and Scikit-learn. The following sub-sections outline the core functional components of the system.

A. Input and Preprocessing Module

This initial module accepts user input in the form of typed or OCR-extracted textual answers. The preprocessing stage cleanses the text by removing stopwords, punctuation, and unnecessary whitespaces. It also includes tokenization and lemmatization to normalize the data. These processes are essential to ensure the semantic consistency of the input before it is analyzed further by the system.

B. Feature Extraction and Similarity Calculation

After preprocessing, the input is transformed into numerical vectors using vectorization methods like TF-IDF and Word2Vec. These vectors are then compared with reference answer vectors using similarity metrics such as Cosine Similarity and Word Mover's Distance (WMD). This step ensures that the system captures not just keyword matches, but also semantic similarities in phrasing and structure.

C. Scoring Using Machine Learning Models

In this module, the extracted features and similarity scores are passed to a trained machine learning model, such as Multinomial Naive Bayes (MNB). The model is trained on a dataset of graded subjective answers, allowing it to predict a score for new responses based on previously learned patterns. Additionally, rule-based adjustments may be applied based on keyword presence or sentence-level match percentages to enhance accuracy.

D. Result Display and Feedback Module

Once a score is generated, the system displays it through a user-friendly graphical interface developed using Python's Tkinter or a web framework like Django. The interface may also provide basic feedback, such as missing key concepts or low semantic relevance, guiding students toward improved answers.

E. Session and Data Management

To enable efficient tracking and management, the system supports session-level history. Each evaluated answer, its predicted score, and the feedback are stored temporarily, allowing users to review past evaluations within the same session. This module plays a vital role in educational analysis and progress tracking over multiple attempts or assessments.

VI. RESULTS

The proposed system for automatic subjective answer evaluation was tested on a dataset consisting of varied descriptive responses across multiple subjects. The evaluation system successfully assessed the answers based on semantic similarity, keyword presence, and contextual relevance, utilizing NLP techniques and machine learning models such as Word2Vec, TF-IDF, Cosine Similarity, Word Mover's Distance (WMD), and Multinomial Naive Bayes (MNB).

During testing, the system demonstrated a high level of accuracy and consistency in grading, with the WMD method outperforming cosine similarity in preserving semantic context. On average, the model achieved a prediction accuracy of approximately **88%** using the MNB classifier. Furthermore, it was observed that the semantic-based scoring yielded results closer to human evaluation when compared with traditional keyword-matching approaches.

The results also showed that the system could efficiently handle noisy or unstructured answers due to the preprocessing techniques implemented (stopword removal, lemmatization, etc.). The interface provided real-time feedback to users, and the stored session history allowed for tracking past evaluations, making the system both practical and user-friendly in academic settings.

In summary, the system achieved its objective of automating subjective answer assessment with reliable accuracy, reducing manual effort, and providing timely evaluation thus enhancing the efficiency and effectiveness of the educational assessment process.

VII. CONCLUSION

This project presents a novel approach for the evaluation of subjective answers using Machine Learning and Natural Language Processing techniques. Two score prediction algorithms were developed to accurately assess student responses. The system utilizes advanced NLP tools such as Word2Vec and Word Mover's Distance (WMD) to maintain the semantic integrity of answers.

Experimental results indicate that the Word2Vec approach consistently outperformed traditional word embedding methods by effectively preserving contextual meaning. Additionally, WMD demonstrated better performance than Cosine Similarity in most cases, contributing to more accurate scoring. These techniques enabled the machine learning model primarily the Multinomial Naive Bayes classifier to deliver an average prediction accuracy of up to 88%.

The proposed solution effectively automates the grading process, reduces evaluation time, and minimizes human bias, while providing consistent and reliable results. With adequate training, the system can independently assess responses without manual intervention, making it a scalable and practical tool for educational institutions.

REFERENCES

[1] J. Wang and Y. Dong, "Measurement of text similarity: A survey," Information, vol. 11, no. 9, p. 421, Aug. 2020.

[2] M. Han, X. Zhang, X. Yuan, J. Jiang, W. Yun, and C. Gao, "A survey on the techniques, applications, and performance of short text semantic similarity," Concurrency and Computation: Practice and Experience, vol. 33, no. 5, Mar. 2021.

[3] M. S. M. Patil and M. S. Patil, "Evaluating student descriptive answers using natural language processing," International Journal of Engineering Research & Technology (IJERT), vol. 3, no. 3, pp. 1716–1718, 2014.

[4] P. Patil, S. Patil, V. Miniyar, and A. Bandal, "Subjective answer evaluation using machine learning," International Journal of Pure and Applied Mathematics, vol. 118, no. 24, pp. 1–13, 2018.

[5] J. Muangprathub, S. Kajornkasirat, and A. Wanichsombat, "Document plagiarism detection using a new concept similarity in formal concept analysis," Journal of Applied Mathematics, vol. 2021, pp. 1–10, Mar. 2021.