Subjective Answer Evaluation Using Machine Learning And Natural Language Processing

Shailaja S. Katti
Computer Science and Engineering
Annasaheb Dange College of Engineering and Technology
Sangli, India
ssk_cse@adcet.in

Yash S. Gawade
Computer Science and Engineering
Annasaheb Dange College of Engineering and Technology
Sangli, India
yashgawade2222@gmail.com

Pratik P. Pattanshetti
Computer Science and Engineering
Annasaheb Dange College of Engineering and Technology
Sangli, India
pratikpattanshetti1007@gmail.com

Gururaj R. Takale
Computer Science and Engineering
Annasaheb Dange College of Engineering and Technology
Sangli, India
gururajtakale0007@gmail.com

Prasad S. Herwade
Computer Science and Engineering
Annasaheb Dange College of Engineering and Technology
Sangli, India
prasadherwade28@gmail.com

***Abstract*—ective answers is a challenging task in education and assessment, often requiring significant time and effort from educators. Auto- mated systems leveraging Machine Learning (ML) and Natural Language Pro- cessing (NLP) offer promising solutions to this challenge. In this project, we propose a novel approach to subjective answer evaluation, utilizing advanced ML and NLP techniques.The primary objective of our project is to develop a system capable of accurately assessing subjective answers provided by students in educational assessments. We aim to create a tool that can mimic human-like evaluation, providing insightful feedback to both students and educators.Furthermore, we employ NLP techniques to extract meaningful features from textual answers. These features will enhance the model's understanding of the content and context of the answers, enabling it to make more informed evaluations. Overall, our project aims to contribute to the advancement of automated subjective answer evaluation systems, providing valuable tools for educators to streamline the assessment process and enhance the learning experience for students. Through the integration of ML and NLP techniques, we strive to develop a robust and reliable system capable of delivering accurate and insightful evaluations of subjective answers. In conclusion, our project contributes to the advancement of automated subjective answer evaluation systems, offering a valuable tool for educators to enhance the efficiency and effectiveness of educational assessments. Through the fusion of ML and NLP techniques, we aim to develop a versatile and reliable system that empowers educators and improves the learning experience for students.**

***Index Terms*—Automated subjective answer evaluation, machine learning, score, NLP etc**

I. Introduction

In the current educational landscape, the assessment of subjective answers remains a critical yet challenging component. Traditional methods largely depend on human evaluators who bring their own perceptions and bi- ases, inevitably leading to variations in grading standards. Furthermore, with the expanding scope of education, both in terms of student numbers and edu- cational formats such as online learning, the need for a scalable, consistent, and efficient evaluation system has become more pressing. This project proposes the use of machine learning (ML) and natural language processing (NLP) as tools to revolutionize the way subjective answers are assessed.

This initiative addresses the critical need for a more efficient, un- biased, and scalable assessment process, which is becoming increasingly vital in both conventional and modern educational platforms that are dealing with large volumes of students and digital coursework.As educational institutions continue to grow and diversify, the limitations of human grading—such as inconsistencies, bias, and the sheer time required to evaluate large sets of subjective answers—become more pronounced. This project aims to mitigate these challenges by developing an automated system that can evaluate written responses with the same depth of understanding as a human grader but with greater consistency and efficiency.

Subjective answer evaluation typically involves complex, open-ended responses that require an understanding of context, nuanced language, and the ability to make informed judgments. Historically, this task has been the exclusive domain of human assessors. However, this approach is filled with challenges such as scalability, speed, and cost-efficiency, not to mention the potential for human error and bias. The integration of technology in educa- tional assessment has been gradual but has primarily focused on objective, easily quantifiable

responses due to their straightforward nature in automated assessment.

The advent of advanced computational techniques and algorithms in machine learning and natural language processing presents a new opportunity to tackle the more intricate problem of subjective response evaluation. These technologies can analyze large datasets and understand complex patterns in text, making them ideal for interpreting and assessing written responses.This project aims to explore the fusion of ML and NLP techniques to develop a robust system for evaluating subjective answers.By leveraging the power of machine learning algorithms and natural language understanding, this system seeks to achieve several goals such as automated grading, continuous improve- ment, scalability etc.

II. Literature Review

By concentrating on the terms that both texts share, **Oghbaie and Zanjireh** created a novel method for calculating how similar two documents are to one another. In order to more accurately evaluate textual similarities, their approach—known as the Pair-wise Document Similarity Measure, or PDSM—evolves from the preferred properties technique. This measure has been successfully used in a number of text mining scenarios, such as K-means clustering and document categorization, which are specifically designed for single-label classification tasks. The effectiveness of the PDSM has been shown by improvements in the precision and speed of finding and classifying related documents in sizable datasets. Text mining has advanced significantly with the creation of PDSM, allowing for more complex analyses of textual data. This method supports more documents and increases the scalability of document processing operations.The PDSM has proven especially effective in text mining applications such as K-means clustering and document categorization, particularly within single-label classification tasks. It facilitates more sophisticated textual analyses and supports enhanced scalability in document processing, which is essential in handling the increasing volume of data. This development not only improves the accuracy and speed of document-related tasks but also advances the capabilities of text mining, providing a robust tool for researchers and practitioners dealing with extensive text-based datasets.[1]

Muhammad Farrukh Bashir and Shahab S. Band explored the application of machine learning (ML) algorithms and natural language processing (NLP) techniques to develop an automated system for evaluating subjective answers. By integrating ML and NLP into the assessment process, the researchers hope to streamline grading, reduce the time educators spend on manual review, and enhance the reliability of evaluating students' written work. This advancement in educational technology marks a significant step forward in the adoption of artificial intelligence tools in academic settings, offering a more standardized and efficient method for evaluating complex, open-ended responses.Their study focused on using various computational functions to assess the degree of similarity between pairs of sentences. This approach is critical in understanding how closely a student's response matches a model answer, thus facilitating a more objective and efficient grading process. The researchers employed diverse algorithms that analyze text-based responses, comparing semantic elements and syntactic structures within the answers to gauge similarity. This method not only enhances the reliability of subjective answer assessment but also significantly reduces the time and effort traditionally required in manual grading. The innovative use of ML and NLP in education underscores a transformative step toward integrating artificial intelligence in academic evaluation processes, aiming to achieve higher accuracy and consistency in results.[2]]

In the field of automated essay scoring (AES), significant advancements have been made to enhance the accuracy and efficiency of evaluations. One notable contribution is from a study by **Taghipour and Ng** in 2016, which pioneered the integration of deep learning technologies into AES systems. Their research introduced a novel approach by combining both convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to develop a more robust scoring mechanism. This hybrid model was particularly effective in analyzing and understanding complex semantic features and the contextual nuances of written text. Their findings indicated that such deep learning frameworks could significantly outperform traditional scoring algorithms by adapting more effectively to the variances in linguistic expression found in student essays. This approach not only marked a progressive step towards more accurate assessments of written submissions but also highlighted the potential of neural networks in educational applications.This research not only pushed the boundaries of what's possible in AES by integrating sophisticated neural network architectures but also set a precedent for the future use of deep learning in educational assessments. The success of this model highlighted the potential for neural networks to revolutionize how educational systems evaluate and understand student submissions, paving the way for further innovations in the field. The work of Taghipour and Ng stands as a landmark study, signaling a shift towards more advanced, fair, and reliable scoring systems in education.[3]

Attali and Burstein (2006) investigated the integration of Natural Language Processing (NLP) tools in automated essay scoring (AES), focusing on features such as syntactic complexity and lexical diversity. Their research underscored the potential of these language indicators to enhance the evaluation of written submissions. Further advancements in the field were propelled by the development of sophisticated technologies such as word embeddings and neural networks. This technological progression was highlighted in the work of Mohler et al. (2011) titled "Automated Assessment of Short Text Answers." Their research provided substantial evidence supporting the use

of advanced machine learning techniques to assess short written responses effectively. This body of work collectively illustrates significant strides in leveraging computational methods to improve the objectivity and efficiency of text evaluation in educational settings.These studies collectively underscore the trajectory of research in AES, highlighting how computational methods have increasingly been harnessed to refine the objectivity and efficiency of educational assessments. The integration of sophisticated NLP tools into AES systems promises to offer educators and institutions a more reliable means of evaluating student performance through written tasks. [4]

The research conducted by **Berke Oral, Erdem Emekligil, and Secil Arslan** represents a pioneering effort in the field of information extraction from scanned documents by integrating textual information. This study is notable for its application of auxiliary learning techniques that simultaneously leverage multiple tasks to enhance the primary goal of text extraction. Additionally, it delves into the positioning characteristics of words within the documents, which plays a crucial role in understanding the structure and layout of the text. The research also explores the impact of utilizing diverse word representations on the effectiveness of the information extraction process. By examining these various elements, the study contributes significant insights into the optimization of text extraction technologies, potentially influencing future methodologies in handling scanned documents and similar datasets.The implications of this research are profound, potentially setting new directions for the development of technologies involved in processing and managing scanned documents and other similar data-intensive tasks. By integrating auxiliary learning and examining multiple dimensions of word representation and positioning, the study not only improves current methodologies but also opens up new avenues for future innovations in the field of document analysis and information retrieval. This comprehensive approach not only enhances the primary task of text extraction but also contributes to the broader understanding of document handling technologies.[5]

Study conducted by **Xinming Hu and Huosong Xia**, the focus is on developing an automated evaluation system for subjective questions using latent semantic indexing (LSI). The approach involves creating a term-document matrix from reference answers, which are processed using Chinese automatic text segmentation and subject-specific ontologies. This matrix is subsequently mapped into a lower-dimensional space defined by 'k' dimensions using LSI, a technique built on statistical analysis. The reduction of dimensionality is accomplished through singular value decomposition, a key component of LSI that effectively addresses issues related to synonymy and polysemy in language processing. This methodology underscores the potential of LSI in enhancing the accuracy and reliability of automated systems for assessing subjective responses.The inclusion of LSI in the system underscores its capability to improve the interpretation of nuanced textual data. This innovation is particularly significant in educational settings where subjective assessments are commonplace and demand high accuracy and fairness in evaluation. Hu and Xia's research marks a step forward in the application of advanced computational techniques to the field of automated assessment, showcasing the potential of integrating sophisticated statistical analysis tools like LSI to refine the evaluation process.[6]

Laurie Cutrone and Maiga Chang explore the diverse landscape of learning management systems (LMSs) currently in use. They highlight that while LMSs offer extensive functionalities, they particularly focus on the segment dedicated to student assessment management. This specific function of an LMS is crucial yet presents challenges, especially in its ability to handle different types of evaluations autonomously. The authors point out a significant limitation in the case of open-ended questions, where the system lacks the capability to autonomously assess student responses. Such assessment formats require manual grading by educators, as the LMS is not equipped with the necessary tools to automatically evaluate free-form text or complex answer structures. This underscores a gap in the automated capabilities of current LMS platforms, suggesting a need for advancements in this area to reduce the reliance on human intervention and enhance the overall efficiency and effectiveness of online learning assessments.The enhancement of assessment capabilities in LMSs would also likely improve the learning experience for students, providing quicker and potentially more detailed feedback on their progress. Overall, the work of Cutrone and Chang sheds light on a vital area of need in digital education tools, calling for increased innovation and development to bridge the current gap in LMS functionality. This would not only streamline assessment processes but also contribute to the broader goals of modernizing and improving educational outcomes through technology.[7]

In their detailed research, **Jiapeng Wang and Yihong Dong** dive into the current landscape of similarity measurement, providing an extensive evaluation of the various methodologies used in this field. Their study not only highlights the strengths and weaknesses of existing methods but also introduces a refined classification system for categorizing text similarity measurement algorithms. They define the method of text similarity assessment through two primary dimensions: text distance and text representation. This framework is designed to enhance understanding and guide further research and practical applications in the area of text similarity. The authors also present a vision for the future progression of this field, suggesting directions for upcoming investigations and developments. This comprehensive overview aims to serve as a foundational reference for both scholars and practitioners engaged in the study of text similarity.Through their analysis, Wang and Dong aim to streamline the complexities involved

in selecting appropriate methods for specific applications, thus promoting more efficient and accurate outcomes in practical scenarios. Furthermore, their work paves the way for more targeted research efforts, highlighting gaps and suggesting potential areas for innovation within the field of text similarity.[8]

III. PROPOSED METHODOLOGY

Following diagram shows the system architecture:

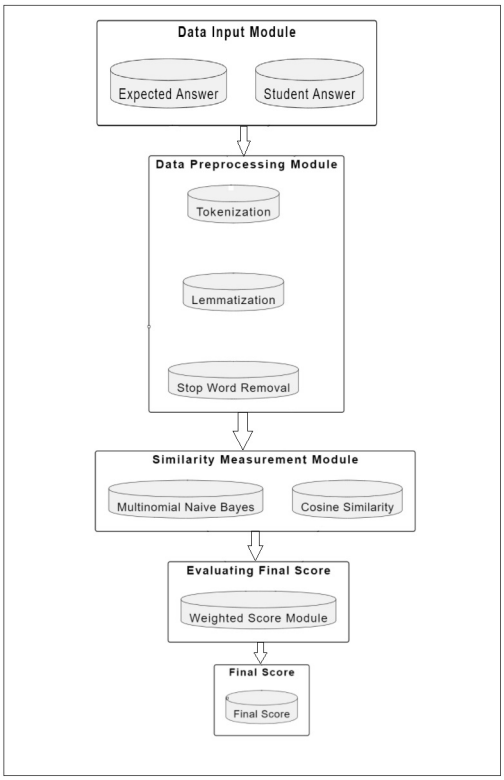

Fig. 1. Flow Diagram

This project aims to revolutionize subjective answer evaluation by using ML and NLP. It begins by addressing the scarcity of labeled subjective question-answer data, creating a corpus sourced from diverse websites and blogs hosting such content.

This project focuses on evaluating subjective answers using ML and NLP methods. It involves creating a labeled corpus from websites, preprocessing inputs, measuring similarity using methods, predicting scores based on matched sentence pairs, employing machine learning models like Multinomial Naive Bayes, and refining scores through validation against predicted classes. Ultimately, it offers a comprehensive approach for assessing subjective answers through a systematic combination of NLP and achine learning methodologies.

1] Preprocessing Module: The preprocessing module is a crucial component of text data analysis, specifically designed to prepare the text for further processes such as similarity measurement and result prediction. This module consists of three fundamental steps that transform raw text into a structured format that can be effectively analyzed.

Responsibility: This module is responsible for preparing the text data before similarity measurement and result prediction. It involves three main preprocessing steps:

- Tokenization: Breaking down the text into individual words or tokens.
- Lemmatization: Reducing words to their base or dictionary form to normalize variations (e.g., ”running” to ”run”).
- Sentiment Analysis: Assessing the sentiment or emotional tone of the text, which can provide additional context for result prediction.

Implementation: Utilizes libraries such as NLTK (Natural Language Toolkit) for tokenization and lemmatization, and sentiment analysis tools.

2] Similarity Measurement Module: The Similarity Measurement Module within this project plays a pivotal role by quantifying the alignment between a student's response and the expected answer. This module leverages multiple similarity metrics, each designed to capture different dimensions of text comparison, providing a comprehensive analysis of how closely the student's answer mirrors the expected one.

Responsibility: Calculates various similarity scores between the expected answer and the student's answer to assess how closely they align. These scores help quantify the similarity between the two texts and provide different perspectives on their similarity.
Similarity Metrics:

- Exact Match: Determines if the student's answer exactly matches the expected answer.
- Partial Match: Measures the degree of overlap between the tokens in the student's answer and the expected answer.
- Cosine Similarity: Computes the cosine of the angle between TF-IDF vectors of the expected and student answers, capturing semantic similarity.
- Semantic Similarity: Calculates similarity based on embedding representations of the text, which captures semantic meaning beyond simple word overlap.
- Coherence Score: Evaluates the coherence of the two texts based on their token overlap and length.
- Relevance Score: Assesses the relevance of the student's answer to the expected answer by comparing common tokens.

Implementation: Utilizes various techniques such as vectorization, word embeddings, and similarity metrics from libraries like scikit-learn and Hugging Face Transformers.

3] Result Prediction Module: The Result Prediction Module plays a pivotal role in determining the accuracy of a student's response by utilizing the similarity scores generated in the Similarity Measurement Module. This module uses advanced machine learning techniques to enhance the predictive accuracy of the assessment system.

Responsibility: Utilizes the similarity scores computed in the Similarity Measurement Module alongside machine learning models (such as Naive Bayes Classifier) to predict the correctness of the student's answer. It combines similarity-based features with other relevant features to make more accurate predictions.

Implementation: Feature Engineering: Extracts features from the similarity scores and potentially additional features like the length of the answers, sentiment scores, etc.

Machine Learning Model (Naive Bayes): Trains a Naive Bayes Classifier using the extracted features. The model learns the relationship between these features and the correctness of the student's answer from a labeled dataset.

Prediction: Uses the trained Naive Bayes model to predict whether the student's answer is correct or incorrect based on the computed features.

4] Weighted Module: The implementation of this weighted average calculation involves assigning different weights to each similarity score based on their significance in determining the overall similarity. These weights are not arbitrarily chosen; they are meticulously configured to reflect the relative importance of each similarity metric in the specific context of subjective answer evaluation. The determination of these weights can be grounded in domain expertise, where a deep understanding of the subject matter helps identify which aspects of similarity are most critical.

Responsibility: Combines the different similarity scores obtained from the Similarity Measurement Module, including those from the Naive Bayes Classifier in the Result Prediction Module, into a single composite score using a weighted average approach.

This composite score provides a comprehensive assessment of the similarity between the expected and student answers, considering the relative importance of each similarity metric.

Implementation: Calculates the weighted average of similarity scores, where weights are assigned based on the relative importance of each metric. These weights can be predefined based on domain knowledge or learned through optimization techniques such as grid search or cross-validation.

IV. Conclusion And Future Work

Conclusion: The subjective answer evaluation project underscores its significant contributions to automated assessment systems leveraging machine learning (ML) and natural language processing (NLP) techniques. Through meticulous implementation and evaluation, the project successfully demonstrated the feasibility and efficacy of employing advanced algorithms to evaluate subjective answers. We've laid the groundwork for streamlined assessment processes, empowering educators with invaluable tools for personalized learning experiences.

Future Work: Building upon the current achievements of the project in subjective answer evaluation using machine learning (ML) and natural language processing (NLP), the future scope of this initiative holds significant promise for even greater enhancements in accuracy and functionality. One key area for future development is the incorporation of more advanced NLP capabilities to enrich the understanding of textual data. Techniques such as deep contextual embedding, which utilize state-of-the-art models like BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer), could allow the system to achieve a more profound comprehension of language nuances, idiomatic expressions, and complex sentence structures that are typical in student responses.

V. Acknowledgment

We would like to extend our heartfelt thanks to everyone who played a role in advancing and bringing to fruition our research paper titled "Subjective Answer Evaluation Using Machine Learning and Natural Language Processing." Special appreciation goes to our advisor, Prof. Shailaja S. Katti from the Department of Computer Science at ADCET, Sangli, whose invaluable advice, support, and encouragement were crucial throughout the research process. We hope that our work contributes to the ongoing initiatives aimed at improving and fostering mentorship relationships within different educational settings, and we are excited about the continued exploration of the possibilities within the Subjective Answer Evaluation Using Machine Learning and Natural Language Processing.

References

[1] M. Oghbaie and M. M. Zanjireh, "Pairwise document similarity measure based on present term set," J. Big Data, vol. 5, no. 1, pp. 1–23, Dec. 2018.

[2] Muhammad Farrukh Bashir, Hamza Arshad, Abdul Rehman Javed, Natalia Kryvinska ,Shahab S. Band."Subjective Answers Evaluation Using Machine Learning and Natural Language Processing",Nov. 2021.

[3] Taghipour and Ng.,"A Deep Learning Approach to Automated Essay Scoring", 2016.

[4] "Automated Assessment of Short Text Answers" by Mohler et al. (2011).

[5] B. Oral, E. Emekligil, S. Arslan, and G. Eryigit, "Information extraction ˘ from text intensive and visually rich banking documents," Inf. Process. Manage., vol. 57, no. 6, Nov. 2020, Art. no. 102361.

[6] X. Hu and H. Xia, "Automated assessment system for subjective questions based on LSI," in Proc. 3rd Int. Symp. Intell. Inf. Technol. Secur. Informat., Apr. 2010, pp. 250–254.

[7] L. A. Cutrone and M. Chang, "Automarking: Automatic assessment of open questions," in Proc. 10th IEEE Int. Conf. Adv. Learn. Technol., Sousse, Tunisia, Jul. 2010, pp. 143–147.

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