International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 08 Issue: 06 | June - 2024 SJIF Rating: 8.448 ISSN: 2582-3930

Evaluating Subjective Answers with Machine Learning and Natural Language Processing

Varsha Pandagre, Deepali Hajare, Rupesh Suryawanshi, Yash Patil, Pratik Pawar , Suvam Kumar Verma
Department of Artificial Intelligence and Data Science
Dr. D .Y. Patil Institute of Engineering, Management and Research
Akurdi, Pune, India
{suryawanshirupesh25, yrajpatil2002, pawarpratik1932, vksuvam}@gmail.com
{varsha.pandagre, deepali.hajare}@dypiemr.ac.in

***Abstract*— Evaluating papers with subjective answers can be both challenging and exhausting when done manually. One of the biggest hurdles in analyzing subjective papers through Artificial Intelligence (AI) is the limited understanding and acceptance of data. Many existing AI approaches for scoring students' answers rely on simplistic metrics such as word counts or specific keywords. Additionally, the dearth of curated datasets for training such AI models further complicates the process. Most previous methods in this domain have utilized predefined solution keys containing expected answers to questions. However, it's crucial to note that these solution keys might themselves contain inaccuracies or contextually incorrect content, leading to sub-optimal evaluations of students' responses. The proposed methodology addresses these challenges by employing a Large Language Model (LLM) to generate answers based on a relevant textbook. This LLM, equipped with the contextual knowledge from the textbook, forms responses to given questions. To assess student answers, we introduce a novel approach involving the comparison of embeddings between the student's response and the LLM-generated answer, utilizing cosine similarity. The resulting similarity score serves as a metric for determining the quality of the student's response. Furthermore, to implement our solution, we leverage the Langchain framework, ensuring a robust and efficient framework for the evaluation process..**

***Keywords*– Subjective Answer Evaluation, Large Language Model (LLM), Cosine Similarity, Machine Learning, Langchain.**

I. INTRODUCTION

In an era marked by digitalization and globalization, the educational sector confronted a multitude of chal-lenges. Teachers found themselves grappling with growing expectations for delivering quality education, personalized learning experiences, and a diverse array of assessment methods. However, a significant portion of their responsi-bilities proved to be tedious, time-consuming, and sus-ceptible to human errors. Notably, the laborious task of comparing students' answers with correct solutions, especially in the context of large-scale or open-ended tests, was a recurring challenge. As depicted in the figure below, the number of test takers for standardized exams like the SAT and ACT continued to rise steadily. This trend was mirrored across various tests. The reliance on human evaluators became increasingly impractical due to the growing volume of assessments, contributing to a time-intensive process.

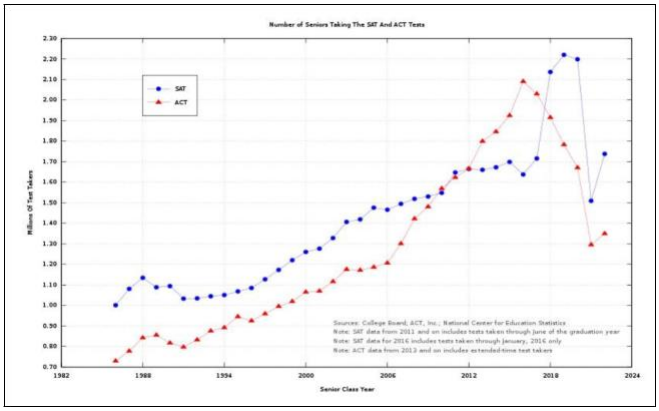

Fig. 1. Number of students taking the SAT and ACT tests [7].

To address these pressing issues, it became impera-tive to explore innovative solutions. The introduction of machine-based evaluation represented a significant leap forward for the education sector. Leveraging natural lan-guage processing and machine learning techniques, these systems could automatically assess the semantic similarity between texts and assign scores based on predefined criteria. This transformative shift relieved educators from the burdens of manual labor in routine tasks such as answer comparisons, enabling them to allocate more time to their core responsibilities.

Furthermore, machine-based evaluation significantly enhanced the quality and transparency of the assess-ment process. Machines provided consistent and objec-tive feedback to both students and teachers, free from the influence of human biases or emotions. Additionally, they generated detailed reports and analytics, empowering educators to monitor student progress and performance. This, in turn, allowed for the identification of students' strengths and weaknesses, facilitating the design of more tailored curricula that catered to their specific needs and interests.

The integration of AI into the subjective answer evalu-

ation process presents a promising solution, yet it faces obstacles related to data comprehension and the scarcity of suitable training datasets. Many existing AI approaches rely on basic word counts or specific keywords, limiting their effectiveness. The proposed methodology adopts LLMs, utilizing a relevant textbook as a knowledge source for generating answers to questions. The evaluation of a student's response involves comparing embeddings of the student's answer and the LLM-generated answer using cosine similarity, determining the quality of the response. This approach leverages advanced language models and aligns them with a reference textbook, promising a more structured and accurate method for assessing student answers. The integration of langchain further enhances the capabilities of this methodology, potentially revolu-tionizing subjective paper evaluation.

II. LITERATURE SURVEY

A. Related Work

In a related study by Farrukh Bashir et al.[1], the focus centers on evaluation of subjective answers using machine learning and natural language processing techniques. It emphasizes the challenges in assessing subjective answers due to their varied lengths, vocabulary, and the use of synonyms. The proposed approach aims to automate the evaluation process by leveraging tools like Wordnet, Word2vec, Word Mover's Distance (WMD), cosine simi-larity, Multinomial Naive Bayes (MNB), and TF-IDF.The solution involves using solution statements and keywords to predict the grades of answers. The WMD method outperforms cosine similarity in the evaluation process, achieving an accuracy of 60% without MNB. The error rate is further reduced by 1.3% with the inclusion of MNB. The preprocessing module includes steps like tokeniza-tion, stemming, lemmatization, stop words removal, case folding, and synonym attachment to enhance the text data before analysis.

The aim is to streamline the evaluation of subjective an-swers, making it more efficient and less resource-intensive compared to manual assessment.Furthermore, the docu-ment highlights the importance of curated datasets for training and testing the proposed model. It mentions the use of a corpus containing over 1,000 short subjective questions, each with a correct answer and 20 student responses, all annotated with necessary keywords. The preprocessing steps ensure that the text data is trans-formed into a numerical form suitable for machine learn-ing algorithms.

The technical background section categorizes evalu-ation techniques into statistical, information extraction, and full natural language processing methods. It discusses the limitations of existing studies, such as handling syn-onyms, varying answer lengths, and random sentence order. The proposed approach addresses these challenges by incorporating advanced NLP techniques and machine learning models.

The paper also references related works by Kusner et al. and Kim et al., highlighting the use of Word Mover's Dis-tance and lexico-semantic patterns for text analysis. These studies demonstrate the effectiveness of novel approaches in improving text evaluation and classification tasks. The experimentation results show a 5% difference in grading between manual assessment and the proposed automated system.

Shubham kumar Sinha and Sachin Yadav [3] underscore the need for a subjective answer evaluation by focusing on leveraging Natural Language Processing (NLP) techniques for the automated evaluation of student answer scripts. It addresses the challenges associated with manual assess-ment and the advantages of adopting automated methods. The study aims to streamline the grading process by utilizing NLP algorithms, text summarization, and sim-ilarity metrics to score student responses accurately and efficiently. The study proposes a systematic approach that combines NLP-based algorithms with text summarization techniques to evaluate student answers.

The authors employ keyword-based summarization and similarity measures such as cosine similarity and bigram similarity algorithms in their research, the system aims to provide a reliable and objective assessment of subjective responses. This automated evaluation process is designed to reduce the workload on teachers and ensure consis-tent grading standards. The evaluation system incorpo-rates various components, including text preprocessing, keyword extraction, similarity computation, and summa-rization algorithms. These components work together to analyze student responses, identify key phrases, calculate similarity scores, and generate concise summaries of the answers. By automating these processes, the system can efficiently evaluate a large number of answer scripts with high accuracy. By combining text summarization, similar-ity metrics, and NLP algorithms, the system offers a robust solution for grading subjective responses. The research highlights the importance of automated assessment in improving grading efficiency, reducing bias, and provid-ing valuable feedback to students. Overall, the research contributes to the advancement of automated evaluation systems in educational contexts.

P. S. Devi, S. Sarkar, T. S. Singh, L. D. Sharma, C. Pankaj and K. R. Singh [5] ,An Approach to Evaluating Subjective Answers using BERT model" discusses the challenges of evaluating subjective answers in the education sector, particularly in the context of online assessment and the limitations of existing Natural Language Processing (NLP) methods. It introduces the Bidirectional Encoder Repre-sentation Transformers (BERT) model as a state-of-the-art method for understanding language and predicting the next sentence. The proposed methodology aims to evaluate subjective answers with semantic meaning using BERT for word embedding and sentence conversion into vector space using pooling methods. The system utilizes a user interface to gather questions and answers for storing in a data storage system, employing the BERT model

refined using deep learning and mathematical methods. The results of the proposed system demonstrate its ability to distinguish semantically related answers with the target answer, irrespective of the length of the answers, providing a systematic approach for grading subjective answers using similarity percentage.

Key quotes: - "In this work, a mathematical method is proposed for evaluating subjective answers using Bidi-rectional Encoder Representation Transformers for word embedding and convert the sentence into vector space using pooling method for representing similar sentences."

- "The BERT model is used with machine learning methods to transform the sentence into vector space. The vector space is used to calculate percentage of similarity. The similarity of the sentences with percentage is ob-served and evaluated."

- "The proposed system is designed to evaluate and check identical and semantic related answers with the target answer by providing the similarity percentage"

In a related study by Piyush Patil [6],the manual system for evaluating subjective answers in technical subjects is time-consuming and requires significant effort from evaluators. Subjective answers are assessed based on pa-rameters like question-specific content and writing style, which can vary in quality and consistency when evaluated by humans. To address this issue, the proposed sys-tem utilizes machine learning (ML) and natural language processing (NLP) techniques to automate the evaluation process.

The algorithm in the system performs tasks such as tokenizing words and sentences, part-of-speech tagging, chunking, chinking, lemmatizing words, and wordnetting to evaluate subjective answers. Additionally, the algorithm provides the semantic meaning of the context in the an-swers. The system is divided into two modules: extracting data from scanned images and categorizing it, and using a classifier to assign marks to the answers based on training data. The marks assigned by the classifier represent the final output of the evaluation process.

The main aim of the project is to ensure user-friendly and more interactive software to the user. The online eval-uation is a much faster and clear method to define all the relevant marking schemes. It brings much transparency to the present method of answer checking The answers to all the questions after the extraction would be stored in a database. The database is designed as such that it is very easily accessible. Automating repetitive tasks has been the main aim of the industrial and technological revolution.

III. METHODOLOGY

A. ML Model Framework

The model framework, initiating with phases such as 1) Data Selection through Dataset, 2) Data pre-processing,
3) Data transformation using ML models, and 4) Feature selection.

B. System Architecture

The system architecture involves collecting data, pre-processing it, training and testing machine learning algo-rithms, evaluating their performance. System model can be implemented using software tools and technologies, such as python. Figure 2 shows the system architecture for the proposed model. The system architecture com-prises several interconnected components to facilitate an efficient and contextually-aware assessment process. At its core, the system integrates access to relevant textbooks, leveraging them to contextualize student responses and assign scores accordingly. This contextualization ensures alignment with the syllabus and educational standards.

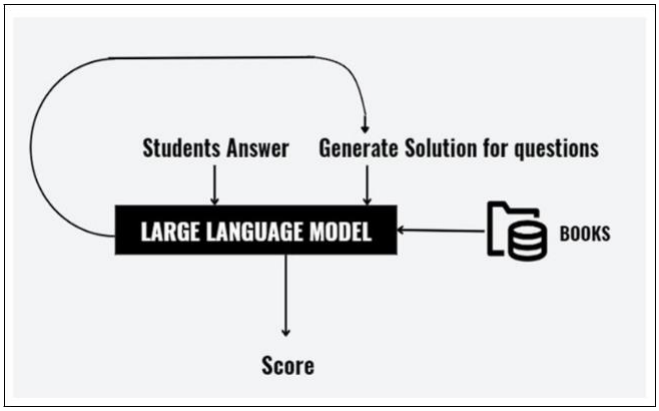

Fig. 2. Flow chart of the result prediction module.

a. Textbook Integration:

The integration of textbooks into the system is pivotal for ensuring access to pertinent content. This process begins with the construction of a comprehensive textbook database, a repository meticulously curated to accommo-date a vast array of subjects and disciplines. Through sophisticated frameworks like ChromaDB, Langchain, or EmbedChain, textbooks undergo a transformative journey into a vectorized format. This conversion is not merely a mechanical process but a strategic endeavor aimed at imbuing textual information with structured vector repre-sentations. These representations serve as the backbone for efficient retrieval and analysis, enabling the system to sift through voluminous data and extract contextually relevant content with precision. Each framework brings its unique set of capabilities to the table, whether it's ChromaDB's focus on semantic coherence or Langchain's prowess in syntactical analysis. Through this integration, the system transcends mere aggregation, evolving into a dynamic platform that empowers users to delve deep into the intricacies of their chosen subjects. Consequently, learners and educators alike benefit from a wealth of curated knowledge, accessible at their fingertips. Such integration underscores the symbiotic relationship be-tween technology and education, fostering a harmonious blend where the digital realm complements and enhances traditional learning paradigms.

b. Input Processing:

The system uses OCR tech to extract text from the uploaded question pdf and answer pdf. That is questions and answers input. Once the system receives inputs from the user, encompassing both the student's response and the associated question, it triggers a multi-step processing pipeline designed for comprehensive evaluation. The first step involves directing the question through a sophisti-cated Language Model (LM) specially calibrated for this purpose. This LM meticulously sifts through the vast textbook database to retrieve information relevant to the query at hand. The retrieved context acts as a cornerstone for subsequent assessments, furnishing the system with a contextual framework upon which further analysis can be built. This initial processing stage serves to lay the groundwork for the system's understanding of the ques-tion and the broader topic it pertains to. By harnessing the power of natural language processing and machine learning, the system navigates through the intricacies of textual data to glean insights essential for accurate assessment. This approach ensures that the subsequent steps in the pipeline are rooted in a solid foundation of contextual understanding, thereby enhancing the accuracy and relevance of the evaluation process.

c. Language Model Integration:

At the core of the system's functionality is its seamless integration with a Language Model (LM), a cutting-edge artificial intelligence module designed to comprehend and generate human-like text. By harnessing the contextual insights gleaned from the textbooks, the LM acts as a powerful engine, generating responses to user queries. Through a nuanced understanding of the provided ques-tion and the broader context encapsulated in the text-books, the LM crafts responses that are not only relevant but also accurate. This integration allows the system to dynamically adapt to diverse inquiries, offering tailored solutions that reflect a deep understanding of the subject matter. As users engage with the system, they benefit from responses that mimic human thought processes, fostering a more immersive and interactive learning experience. In essence, the LM serves as the bridge between user queries and the wealth of knowledge encapsulated within the textbooks, enabling seamless access to information in a manner that is both intuitive and insightful.

d. Embedding Conversion:

In the embedding conversion process, both the stu-dent's input and the response generated by the language model are transformed into embedding representations. These embeddings are designed to encapsulate not just the surface-level text but also the underlying semantic similarities and subtle nuances inherent in the textual data. By converting textual information into embeddings, the system enhances its ability to perform robust compar-isons and conduct in-depth analyses. This transformation enables the system to recognize and comprehend contex-tual intricacies, thereby facilitating more accurate assess-ments and insightful feedback. Through the utilization of embeddings, the system elevates its capacity to discern and evaluate the richness of language, fostering a deeper understanding of the content at hand.

e. Similarity Calculation:

Formula for cosine similarity between two vectors A and B is:

$$\text{cosine similarity}(A, B) = \frac{A \cdot B}{\|A\| \, \|B\|} \quad (1)$$

- A and B are the two vectors being compared.
- AB denotes the dot product of the two vectors.
- A and B represent the Euclidean norms (magnitude) of the vectors A and B,respectively.

Following embedding conversion, the system computes the cosine similarity between the student's response and the LM-generated answer. This similarity metric quantifies the degree of resemblance between the two texts, provid-ing a measure of alignment and correctness.

f. Score Assignment:

The computed cosine similarity score is then scaled to the relevant grading scale. This scaling is achieved by multiplying the similarity score by the maximum marks assignable for the given question. By aligning the score with the grading criteria, the system ensures fairness and consistency in evaluation.

IV. RESULTS AND DISCUSSION

After conducting the system's evaluation on 20 question-answer pairs, we obtained predictions for the marks attributed by the system to each answer. Addi-tionally, human tutors assigned scores to these answers, considering the same context and question prompts. A successful system would exhibit a strong correlation be-tween its predictions and the scores assigned by the tutors, indicating its efficacy. Our system achieved a correlation score of 0.78 for these samples, demonstrating its potential for practical implementation.

To enhance the system's performance further, we pro-pose fine-tuning the RAG system using curated datasets and employing various methodologies. Additionally, inte-grating diagram processing components into the system could mitigate existing bottlenecks and improve its overall functionality. This avenue remains open for future explo-ration and refinement.

V. CONCLUSION

The proposed methodology presents a novel approach to evaluating subjective answers using machine learning and natural language processing techniques. By harness-ing the power of Large Language Models and advanced NLP algorithms, the process of assessing subjective an-swers can be automated with unprecedented accuracy and efficiency. Moreover, the utilization of the Langchain framework ensures a robust and scalable implementation, suitable for deployment in educational settings. Moving

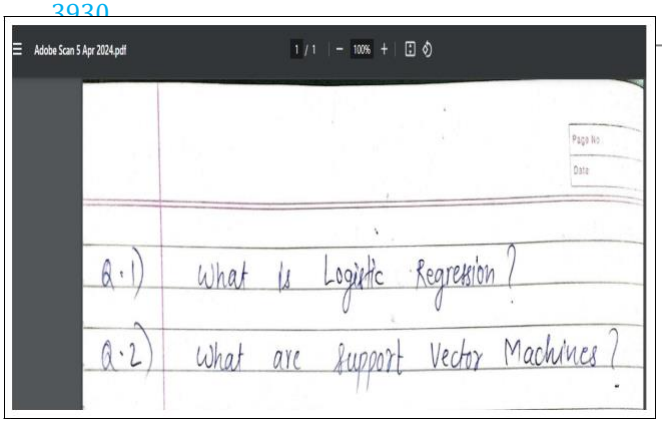

Fig. 3. Questions.

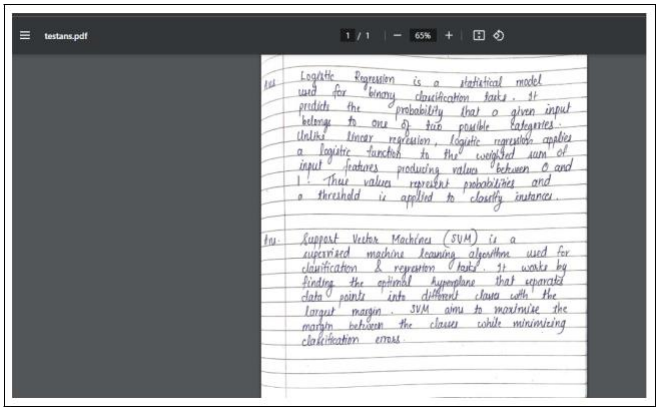

Fig. 4. Answers.

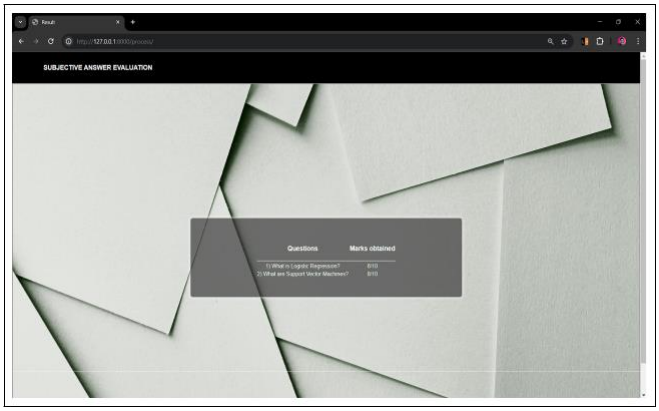

Fig. 5. Questions and their obtained marks.

forward, further research and experimentation are war-ranted to explore additional refinements and extensions to the proposed methodology, paving the way for future advancements in the field of AI-driven education.

REFERENCES

[1] M. F. Bashir, H. Arshad, A. R. Javed, N. Kryvinska and S. S. Band, "Subjective Answers Evaluation Using Machine Learning and Natural Language Processing," in IEEE Access, vol. 9, pp. 158972-158983, 2021, doi: 10.1109/ACCESS.2021.3130902.

[2] Kumari, T.S., Kumar, H., Ahmed, S., Sree, J. and Sravani, J., 2022. Evaluating Descriptive Answers Using Machine Learning And Nat-ural Language Processing.

[3] S. K. Sinha, S. Yadav and B. Verma, "NLP-based Automatic An-swer Evaluation," 2022 6th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 2022, pp. 807-811, doi: 10.1109/ICCMC53470.2022.9754052.

[4] P. S. Devi, S. Sarkar, T. S. Singh, L. D. Sharma, C. Pankaj and K. R. Singh, "An Approach to Evaluating Subjective Answers using BERT model," 2022 IEEE International Conference on Electronics, Com-puting and Communication Technologies (CONECCT), Bangalore, India, 2022, pp. 1-4, doi: 10.1109/CONECCT55679.2022.9865706.

[5] Patil, P., Patil, S., Miniyar, V., Bandal, A. (2018). Subjective Answer Evaluation Using Machine Learning. International Journal of Pure and Applied Mathematics, 23 May

[6] Historical Number of SAT and ACT Test Takers. (n.d.). In Wikimedia Commons. Retrieved April 22, 2024, from https://upload.wikimedia.org/wikipedia/commons/a/a0/Historical_Number_of_SAT_and_ACT_Test_Takers.svg