Journal of Information Systems Engineering and Management
2025, 10(30s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Voice Based Answer Evaluation System for Physically Disabled using Natural Language Processing

Anjusha Pimpalshende[1], Sai Nivas Peddineni[2], Venu Srirama[3], Deepak Thalla[4], Indumathi Katukuri[5]
[1,2,3,4,5]*Department of Computer Science and Engineering, VNR Vignana Jyothi Institute of Engineering and Technology, Hyderabad, India*
[1]*anjusha_p@vnrvjiet.in*, [2]*sainivaspeddineni@gmail.com*, [3]*sriramavenu09@gmail.com*
[4]*deepakthalla123@gmail.com*, [5]*katukuriindumathi2@gmail.com*

ARTICLE INFO

Received: 29 Dec 2024

Revised: 15 Feb 2025

Accepted: 24 Feb 2025

ABSTRACT

The proposed work involves the selection of a subject and evaluation of student responses via a voice-based answer evaluation system that utilizes Natural Language Processing (NLP). This system aims to assist physically disabled individuals, who find it challenging to write their answers by hand. Traditional evaluation methods may become time-consuming, biased, and inconsistent in grading. The approach processes spoken answers, converts the voice signal to text, and finds the relevance of these text answers, according to certain criteria based on a predefined marking scheme. Using NLP techniques ensures maximum grading accuracy with minimum human interaction. The developed system seems to hold promise in accurate evaluation of responses, reduced bias, and greater accessibility for disabled persons.

Keywords: Sentiment analysis, Word embedding, Natural Language Processing, Text-to-Text Transfer Transformer, Latent Dirichlet Allocation, Rhetorical Structure Theory

I. Introduction

Our approach Voice Based Answer Evaluation System is designed to assist physically impaired students who can't write answers in person. Conventional methods of assessing students require written responses that limit the participation of these students in conventionally held examinations. Manual assessors who evaluate answers are commonly upon widely used methods subjecting them to inefficiencies like time-consuming, subjective biases, and variable grading consistency or accuracy for instance, any given answer can be interpreted differently by two or more examiners, and that can lead to forming a disparate assessment score. More so, the grading of voice-inputted assessment still remains human, that is the teacher has to listen, transcribe, and grade the answer increasing the waiting time and chances of error.

To solve these problems, the present system evaluates the spoken answers using Natural Language Processing (NLP) techniques with maximum accuracy. The system takes automatic speech recognition (ASR) as input, capturing the students' voice responses before converting them into text that undergoes deep analysis. The resultant text is analyzed using a number of semantic and syntactic analysis techniques for relevance, correctness, and completeness according to some marking scheme or rule. Text similarity measures such as cosine similarity provision are used so that marking is done according to conceptual correctness. More importantly, T5, LDA (Latent Dirichlet Allocation), and RST(Rhetorical Structure Theory) enable a higher comprehension of the content since they mark the product fairly and consistently under minimal human intervention.

Several existing automated evaluation systems, such as Automated Essay Scoring (AES) systems like e-Rater, IntelliMetric, and Project Essay Grade (PEG), have been developed to assess text responses. Similarly, there are short-answer grading assessment systems like AutoTutor and CoPresent that evaluate based on predefined scoring rubrics. However, all of these systems deal primarily with *typed or handwritten text, which makes them less capable when it comes to evaluations of speech-based answers. The proposed system aims to bridge this gap by incorporating voice recognition with NLP-based automated grading, thereby expanding the scope to include students with physical disabilities.

The effectiveness of such a system is assessed using important performance indicators, including precision, recall evaluating system command, F1, and execution time, thus ensuring that the automated evaluation is basically accurate and happens quite fast. Results show that it really lessens bias, increases grading fairness, and accounts for a better accessibility mode, which sets it strongly for inclusivity in education. The proposed method, eliminating grading restrictions and shortcomings existing in current text-based assessment systems, introduces a unique speech-based solution for answer marking for fair and unbiased evaluation for all students.

II. Literature Survey

In the modern-day world of technology, the ability to convert spoken words into written text, natural language processing and speech recognition have led to the development of systems that can evaluate answers remotely through voice commands which in this case is advantageous to students with disabilities. In most scenarios where examinations are administered, physical examination is routinely followed; this practice however ignores or is prejudicial to students with physical impairments, visual impairment or learning difficulties. To this end, such systems are robust as they enable analysis of oral responses and scoring within a short time. This literature survey covers the important papers, and research done to make such systems focused on their methods, results, conclusions and how they have made education accessible to various clientele.

In the 2018 paper by Md. Motiur Rahman and Fazlul Hasan Siddiqui[1] proposed a system which makes use of Natural Language Processing (NLP) to evaluate answer scripts. This strategy applies NLP techniques to comprehend and assess various written answers in terms of structure, meaning and relevance to the query. The purpose of the system is to provide a more uniform and fair interpretation than the manual grading techniques by recognizing essential concepts and assessing the obviousness of answers offered by students. It also demonstrates how the grading processes in the educational setting are capable of being improved through the use of NLP.

In the paper given by Chahat Sharma, Akash Bishnoi, Akshay Kr. Sachan, and Aman Verma [2], they present a novel essay evaluation system by automated means and with the aid of sophisticated techniques in Natural Language Processing. Quality of an essay covers a diverse number of aspects, including grammar, coherence, structure, and content pertinence which all play a vital role in essay composition. The use of machine learning and NLP algorithms presented aims to offer an objective, clear and fast manner of feedback which would help in the scaling up of the possible ways of evaluating essays.

The article by Chun Then Lim, Chih How Bong, Wee Sian Wong, and Nung Kion Lee,[3] (2021), can be characterized as a systematic review of the existing advances in the area of AES systems. The review examines various designs and frameworks that have been proposed for the purpose of automating scoring of essays in great detail including their pros and cons including challenges faced in terms of specifying their effectiveness and accuracy. In addition, the review also discusses the history of the development of such systems, tracing its roots in the early days when simple statistics were used to even present machine learning techniques. In particular, the review highlights the possible future development and implementation of AES in educational context.

The article authored by Hossam Magdy Balaha and Mahmoud M. Saafan[4] in 2021 comes up with an AECF aimed at grading multiple choice questions (MCQs) and essays, as well as equations matching tasks. The framework incorporates artificial intelligence techniques in order to streamline the evaluation for exams to increase the precision of results and reduce the duration taken in grading. It takes care of various types of question types guaranteeing flexibility in education assessment. The system seeks to improve the effectiveness and expand the scope of computer based examination correction.

In the year 2021, Muhammad Farrukh Bashir, Hamza Arshad, Abdul Rehman Javed, Natalia Kryvinska and Shahab S. Band [5] seek to provide a framework that is based principally on machine learning and natural language processing for effective evaluation of subjective answers. From previous studies, it was noted that traditional automated scoring systems have difficulty scoring subjective answers due to open-ended questions and varying the language. In this paper, we developed a multidisciplinary method that effectively addressed the problem by integrating machine learning and natural language processing for qualitative assessment of the content structure and relevance of the students' responses. The above approach can enhance the accuracy and reliability of the evaluation of subjective answers, which may have implications in educational assessment systems where subjective grading is used.

In 2022, Mayur Wankhade, Annavarapu Chandra Sekhara Rao, and Chaitanya Kulkarni [6] published a paper that reviews numerous sentiment analysis methods. The authors investigate various sentiment analysis techniques, including machine learning, lexicon-based, and hybrid methods, assessing textual affect and opinion these methods employ. They note different cases of sentiment analysis such as social media, product reviews or public opinion and trends, and speak about problems which include - but are not limited to - sarcasm, languages and their variations, and language change over time. The paper describes the present state of affairs and proposes further developments which aim to enhance the scope and effectiveness of sentiment analysis in practice.

Mostafa R. Kaseb, Esraa Rslan, Rasha M. Badry, and Mostafa Ali[7] (2023) present a system for evaluation of short answer questions developed for the Arabic environment. The paper explains the issues associated with the processing and evaluation of textual content in Arabic, including the non-linear nature of the language and the infinities of dialects. The authors specify how these challenges are addressed in practice, in particular using natural language processing technologies. The paper provides performance measures of the system under consideration and discusses potential areas of its use in educational settings.

In the work of Prabakaran N., Kannadasan R., Krishnamoorthy A., and Vijay Kakani [8] (2023) a fully automated script evaluation system is proposed. The system employs a combination of Bidirectional LSTM networks (Bi-) and keyword based pattern recognition algorithms to evaluate the quality of responses. The incorporation of Bi-LSTM for contextual reasoning and keywords for pinpointing specific elements is the essence of the designed system that aims at quality evaluation of written scripts in the shortest time possible and with utmost precision, especially descriptive or open-ended answers. The method therefore improves the effectiveness and ease with which computerized assessment processes can be achieved.

Samuel Tobler[9] in the year 2024 introduces that an advanced artificial intelligence system has been integrated into the process of evaluating the learners' responses to make it more efficient. The technology incorporates generative AI models in the assessment of responses based on content knowledge parameters to enhance fairness and accuracy in assessments. The solution also tackles some of the challenges associated with automatic grading systems, by considering other aspects of the instructional materials being evaluated. The author also discusses artificial intelligence based evaluation systems and their current state in education, as well as their advantages and disadvantages.

The purpose of this paper by Wenbo Xu, Rohana Mahmud, and Wai Lam Hoo[10] (2024) is to assess the feasibility and dependability of applying Automated Essay Scoring (AES) systems at schools and colleges level. The review evaluated systematically the evidence of the effectiveness and challenges of implementing an AES observer in practice. Issues like scoring precision, concern of evaluation bias, and the capability of the systems to give useful feedback were explained. Moreover, the paper looks at the concerns of educators and ways to enhance the real-world application of AES systems.

Lalitha Manasa Chandrapati and Ch. Koteswara Rao[11] (2024) explores the applicability of natural language processing (NLP) paradigms in the automatic assessment of descriptive answers in tests. The authors propose a system for marking open-ended response-type questions based on NLP models which can identify and evaluate the contextual relevance and quality of the content. This idea is driven by the need to reduce the stress involved in the traditional way of grading constructed response-type questions manually and to improve the effectiveness and reliability of the grading of such types of questions.

In the 2024 paper written by Christian Grévisse,[12] the author reviews how large language models (LLMs) can be used to grade short answer questions in medical education. The research pays attention to how large language models (LLMs), which are good with language and context, can analyze medical students' responses not only for correctness but for accuracy, relevance, and depth of the content. This approach answers the concerns about the peculiar range of students that medical education trains whereby understanding is always to the degree of particulars and complexities. Since the grading will be LLM-based, the structure presents a cost effective and fast way of grading students' responses as compared to manual grading which may also help improve the speed and quality of feedback given to the medical students.

In 2024, a novel framework for automatic evaluation of answer scripts or answer sheets was presented by Mohanraj G, Nadesh R. K., Marimuthu M. and Sathyapriya V. [13] This system relies on natural language processing and deep

learning to assess written responses and score them. Given the system's boundaries, improving algorithms applied to grading of more advanced and open-ended answers is one of the goals as it fuses knowledge of patterns in languages through natural language processing and deep learning which is concerned with contextual information. This framework also proposes solutions to existing challenges in computerized grading systems including issues such as variation in the style of writing and the quality of responses offered to questions which can be useful in educational and assessment systems.

According to the 2019 paper published by Harneet Kaur Janda, Atish Pawar, Shan Du and Vijay Mago,[14] essay grading is looked at in a more complicated way than before. It is noted that evaluation of essays should not be limited to their structural (syntax and grammar) analysis and should include their meaning ('semantics') and emotional content ('sentiment') if they are to develop reliable assessment systems based on automation technologies. The research introduces these three aspects to evaluate the writing quality in a more comprehensive way. This strategy seeks to bring automated scoring in line with the grading done by raters, thus improving accuracy of grading educational tests and also making the analysis deeper.

Research Gap:

The present conventional works are entirely based on automated text answer evaluation through the applications of natural language processing (NLP)-AES and deep learning. Most of the works do not explore any attempts toward voice-based assessment in physically handicapped people. Most of the research work does not consider voice recognition combined with NLP for spoken answers. Subjective answer evaluation can not be so easily performed due to language variation and emotional attitudes toward the answers. Domains such as medical or Arabic grading are being discussed, but nothing has been done in the direction of the specific need or framework for differently able people. Further, the immediacy of feedback is not known in those systems, nor does supplemental adaptive learning exist in them. In particular, these highlighted deficiencies need to be filled up by future studies attempting from the point of developing a multimodal voice-based answer evaluation system for hearing-impaired people through integrating speech recognition, NLP, and AI techniques.

Key Findings:

This literature survey reveals that while research until now has focused mainly on text-based answer evaluation using NLP, AES, and deep learning, not much has been targeted toward voice-based evaluation, thus rendering the existing systems less accessible for physically challenged students. The subjective evaluation of answers seems to be a hurdle as variations come in answer structure, coherence, and sentiment, while very little research has been done on multimodal approaches that interface speech recognition with NLP. A small number of studies, including those concerning Arabic and medical fields, have been carried out, but not one has ever offered a general framework for grading that could be adapted to any subject and any disability. Moreover, up-to-date automated grading systems lack any mechanism for real-time feedback or adaptive learning. Recent studies have suggested AI and LLMs as solutions for better evaluation; however, these studies so far have kept intact the spoken response assessment underdeveloped. All these assert the unmet need for an exhaustive voice-based answer evaluation system capable of establishing seamless integration among speech recognition, NLP, and deep learning for enhanced accessibility and accuracy.

III. Proposed Methodology

The process of evaluating descriptive answers is often subjective, time-consuming, and prone to inconsistencies due to human bias. Assessors assign marks not in harmony on the basis of his discretion manual checking of such type of evaluation. The said system introduces measures for the automatic evaluation of descriptive answers provided through voice input using Natural Language Processing (NLP) ,Latent Dirichlet Allocation (LDA), T5 Transformer, and Rhetorical Structure Theory (RST) to evaluate the answers and the system promises to be institutional, consistent, and effective in evaluating responses.

Evaluation would, therefore, be grounded on content relevance, coherence, grammatical accurateness, and structural integrity. The LDA model pulls up significant topics from the answer for assessing correspondence to the expected response, while the T5 Transformer summarizes and grades completeness. Meanwhile, RST analyzes the logical and rhetorical structure of the response so that arguments are properly formed and well supported. Then summing these

models in a weighted aggregation gives the scores for the final grading, which makes the measures fair and transparent regarding evaluations.

Especially useful in teaching settings is this automated measure, which can help school teachers evaluate large numbers of student responses. This, too, allows for instant feedback for students for fine-tuning their articulation and answer formation skills over time.

Module-wise, our proposed work bifurcates itself into two different types: namely text-based and voice-based. Both modules share the same evaluation scheme, which maintains uniformity with respect to evaluation. The primary point of divergence is the methodology of collecting answers-

1.Text-Based Module: In this module, users are required to enter their answers in writing. These textual answers are processed, evaluated, and scored according to predefined criteria.

2.Voice-Based Module: Here, the voice-based system is more empowering for users who may wish or need to give answers verbally, such as physically disabled students who may find typing difficult. Answers are given orally and captured by a speech recognition system that accurately converts spoken words into text. The generated text copies the parameterization for scoring giving the same evaluation again and thus maintaining the uniform pattern.

Hence, by incorporating both modules, our system promotes an inclusive and flexible evaluation approach that allows users with multiple needs while ensuring fair and accurate evaluation.

A. Collection of Data

Collecting the data in the form of google forms from students analyzing them for the evaluation process. Below is the sample list of a dataset:

TABLE:Sample list of dataset

Qid	Question	Sid	Student_Answer	Ideal_Answer
1	What are the major challenges faced in data mining?	s1	As datasets grow in size and complexity, traditional data mining techniques struggle to scale efficiently. Advanced methods are required to handle high-dimensional spaces effectively.	One of the major challenges in data mining is handling huge and complex datasets and dealing with data quality and preprocessing. This involves cleaning the data and removing noise and missing values. It also includes ensuring confidentiality and privacy of data, while getting useful insights from it. Finding the right algorithm for the different datasets, as well as interpreting the meaningful patterns, could also pose some challenges. Another challenge would be the computational efficiency
		s2	Many machine learning and data mining models generate complex patterns that are difficult to interpret. Ensuring that insights are understandable and actionable remains a challenge.	

				and scalability in dealing with the mind-boggling amounts of data.
2	Differentiate between data mining and data warehousing.	s1	A data warehouse provides raw, structured data that can be queried, whereas data mining provides actionable insights, predictions, and relationships between variables.	Data mining is analyzing large sets of data to find patterns, trends, and insights while data warehouse refers to storing and managing structured data from multiple sources so that it could be recalled and analyzed efficiently. Data mining keeps extracting useful information from saved data while data warehousing organizes and consolidates data to provide better access.
		s2	Data warehousing employs ETL (Extract, Transform, Load) processes, while data mining uses machine learning, statistical analysis, and artificial intelligence to uncover patterns.	
3	What is data preprocessing, and why is it necessary in data mining?	s1	Data preprocessing involves detecting and handling noisy or inconsistent data, which can otherwise lead to misleading results and affect the overall performance of data mining models	Data preprocessing involves cleaning, transforming, and preparing raw data to improve the data's quality for the purpose of analysis. In data mining, it is required in order to handle missing values, remove noise, and extract patterns in a manner that is accurate, efficient, and meaningful.
		s2	Preprocessing organizes and cleans data, ensuring it is accurate and ready for use in mining algorithms. This involves resolving missing values and removing inconsistencies.	
4	Explain the concept of data integration and its importance?	s1	Integration is just combine many files like lab and patient file together.	Data integration is the process of merging various types of data into a common view that can be analyzed. It is important to ensure consistency, to eliminate redundancy, and to do most significant thinking, which is usually based on the information that is complete and accurate.
		s2	Data integration combines data from multiple systems, such as integrating patient records with pharmacy data, to create a single, unified source for analysis.	

Our dataset consists of a single chosen subject, for instance, Data Mining, consisting of 50 questions each with its ideal answer; for every question, there are maybe 25–30 student responses. Each response is judged by a human evaluator who marks it, which serves as the output or label. This dataset would be used to train a model using LDA and T5 over a disparate dataset. The trained model can therefore be used to evaluate any student's answer.

A. Ideal Answer:

An ideal answer is a perfect kind of answer that is well-guided, adequately described, all-inclusive of points, and all context details. Answers are thus presented to enhance their readability in a clear and well-structured presentation in separate paragraphs or lines. The teacher formulates this answer for marking and serves solely as an important tool for guideline marking.

B.Student Answer:

A student answer is an answer any learner provides for evaluation. The answer may contain some of the key required words and vary in length, depending on the question types and writing styles of students. As student answers tend to use synonyms more often than the ideal answer, some degree of semantic analysis must be employed for the grading process.

B. Data Pre-Processing:

Data preprocessing in Natural Language Processing (NLP) is a cleaning and preparation of raw text data before a data analyst works on it. The key preprocessing techniques with a brief explanation and examples are below:

1. Tokenization- Split text into words or sentences to facilitate analysis.

Example:- "The cat sat on the mat." → "The", "cat", "sat", "on", "the", "mat.",

2. Lowercasing- Allows all text to be in lower-case letters to provide uniformity so that it is not faced with case-end problems.

Example:- "HELLO World" → "hello world".

3. Stopword Removal- Removes the common words among others which do not add much to the meaning.

Example:- "She is going to the park." → "going park".

4. Stemming- It is a process where the original root of the word is reduced through the removal of suffixes, usually leading to a result that lacks dictionary acceptance.

Example:- "running, runs, ran" → "run".

5.Lemmatization- An operation that arrives at the base or dictionary form of words by applying some language-specific rules.

Example:- "running" and "better" yield "run" and "good", respectively.

6. Removing Punctuations- This is where punctuations mark the lines that do not alter text understanding.

Example:- "Hello, how are you?" → "Hello how are you".

7. Part of Speech (POS) Tagging- This indicates a grammatical label for every word, thus providing insight into the structure of a sentence.

Example:- "Dogs bark." → ("Dogs", Noun), ("bark", Verb).

8. Named Entity Recognition (NER)- The identification of proper names such as persons, locations, or organizations.

Example:- "Barack Obama was the US President." → ("Barack Obama", PERSON), ("US", GPE).

9. Text Normalization- To convert/change the abbreviations, slang, and special characters into standard formats.

Example:- "u r gr8" → "you are great".

10. Remove Special Characters- This is where "cleaning" will entail special symbols like @,#,$,% which carry little relevance.

Example:- "@user123 hello!!!" → "hello"

All of these preprocessing are very vital for models that aid in improving accuracy and efficiency.

The proposed system introduces an automation degree within the speech response evaluation sector. In the very beginning, the speech is recognized, whereby a student's spoken answers are converted to text courtesy of React Speech Recognition Library from react. The response in text either transcribed or typed is further evaluated for scoring, Once the text from the spoken response is obtained, it is subjected to preprocessing which involves tokenization, stopword removal, stemming or lemmatization, and normalization to improve text quality. After pre-processing, the system implements a feature extraction process where, utilizing three main techniques: Latent Dirichlet Allocation (LDA) for topic modeling, T5 Transformer for semantic similarity, and Rhetorical Structure Theory (RST) for coherence analysis. A weighted combination of their individual outputs contributes to the final evaluation score.

Latent Dirichlet Allocation (LDA):

Before LDA application, the student answer and the provided reference (ideal) answer are preprocessed for textual cleaning and standardization.

Once cleaned, it goes into a bag-of-words (BoW) representation, which means every unique word in the dataset gets an index number in a dictionary and each document (the student and reference answers) is represented as a list of counts of words

LDA is a probabilistic topic modeling algorithm that helps to extract meaningful key topics and words from any document. It has the assumption that each document has a mixture of topics and each topic has a set of words belonging to it. Here, applying LDA to both the student answers and the ideal answers to extract thematic topics. Hence, the student answer topic distribution is denoted as Stopic_dis, whereas the ideal answer topic distribution is represented as Itopic_dis. These distributions give the extent to which present topics in each answer are expressed.

In order to compute similarity between student's response and ideal answer, we compute cosine similarity on their topic distributions as follows:

$$\text{Cos\_sim}_{\text{lda}} = \frac{S_{topic-dis} \cdot I_{topic-dis}}{||S_{topic-dis}|| \; X \; ||I_{topic-dis}||}$$

Thus,

Stopic_dis is vector representation of student's topic distribution, whereas

Itopic_dis is vector representation of ideal answer's topic distribution.

Thus the final score of topic similarity obtained from LDA is scaled and contributes 30% to the total evaluation score. As the system evaluates answers out of a maximum of 10 marks, the LDA score is multiplied by 3 to fit into this range.

Example:

Now, consider examples where a student answer and an ideal answer are provided for a question on "Data Mining". After the application of LDA topic modeling, we determined the topic distribution of both answers.

Ideal Answer Topic Distribution I_{topic_dis}=[0.5,0.3,0.2]

(50% Topic A, 30% Topic B, 20% Topic C)

Student Answer Topic Distribution S_{topic_dis}=[0.4,0.4,0.2]

(40% Topic A, 40% Topic B, 20% Topic C)

Cosine similarity gives us $cos\_sim_{lda}=0.974$.

Finally, 30 per cent of the LDA similarity score contributes to total evaluation score.

$$LDA\_Score=cos\_sim_{lda}\times 3$$

$$LDA\_Score=0.974\times 3=2.92$$

T5 Transformer(Text-to-Text Transfer Transformer):

T5 is a Text-to-Text Transfer Transformer that measures semantic similarity between a student's answer and the ideal answer. Whereas LDA focuses solely on topic coverage, T5 makes sure that the student's answer means the same as the reference answer, even when the words differ. Every input is treated within a text-to-text framework, where the model produces a similarity score ranging from values near 0 to values near 1. Thus, if the model predicts a score of 0.85, the student's answer is 85% semantically similar to the ideal answer. By weighing semantic similarity with 50% of the total score, it is scaled as:

$$T5\_Score= semantic\_similarity*5$$

For the above example $T5\_Score=0.85*5=4.25$

Example samples

Student Answer 1 (With High Similarity)

Ideal Answer: "Data mining is the process of extracting patterns from large datasets."

Student Answer: The process of discovering patterns in data is called data mining."

T5 output score: 0.9

Scaled Score: 0.9X5=4.5

Student Answer 2 (With Low Similarity):"Data mining requires a computer and storage."

T5 output score: 0.4

Scaled Score:0.4X5-2.0

If $T5\_Score\approx 5$, the student's answer conveys the same meaning as the ideal answer.

If $T5\_Score\approx 0$, the answer is not related with respect to meaning.

Rhetorical Structure Theory(RST):

Analysis of a student's answer in terms of Rhetorical Structure Theory (RST) helps understand how coherent or inconsistent that student's answer is. Coherence shows how well the components of an answer are organized and how logically and meaningfully they flow. It is not like LDA, which checks the adequacy of thematic coverage, and T5, which checks semantic similarity, for it analyzes sentence relationships, the discourse structure, and logical connections between parts of the text in the student's response. The coherence measure considers dependency parsing conducted using SpaCy to determine how words and phrases are linked within the student's answer.

Example:

Student answer: "Data mining is useful as it finds patterns in large datasets."

Dependency parsing output:

(ROOT

(S	S=Sentence or The main Sentence
(NP Data mining)	NP=Noun Phrase
(VP is useful	VP=Verb Phrase
(SBAR because	SBAR=Subordinate Clause
(S it helps find patterns	
(PP in large datasets))))))	PP= Prepositional Phrase

Here for the cause-effect relationship(because) it made the response coherent.

RST defines rhetorical relations like cause-effect, elaboration, contrast, sequence of presentations, and the like showing how well answers are constructed. For strong coherence, logical connectors are used: "because, therefore," and "however"; follow a clear hierarchical structure; and avoid fragmented sentences. On the contrary, sentence types are disjointed with no logical connections, abrupt shifts on topic, and lack of explanation or reason.

Finally, it extracts dependency relations from the parsed answer and evaluates how many unique discourse relations are realized.

Above all, the Coherence Score is calculated as:

$$\text{Coherence\_Score}=\frac{Number\ of\ unique\ rhetorical\ relations}{Total\ possible\ relations}$$

Coherence is responsible for 20% of the final total score, which gives us scaling as

$$\text{RST\_Score}=\text{Coherence\_Score}\times 2$$

Final Score is calculated using the formula:

Final_Score=(LDA_Score×3)+(T5_Score×5)+(RST_Score×2). All scores are normalized between 0 and 1 and the final score calculated out of 10.

Example Calculation

The table below shows the individual scores for LDA (topic similarity), T5 (semantic similarity), and RST (coherence), along with their weighted contributions to the final evaluation score.

Metric	Score (out of 1)	Weighted Contribution
LDA Score	0.8	0.8×3=2.4
T5 Score	0.7	0.7×5=3.5
RST Score	0.9	0.9×2=1.8
Total Score	7.7/10	

System Architecture of proposed system:

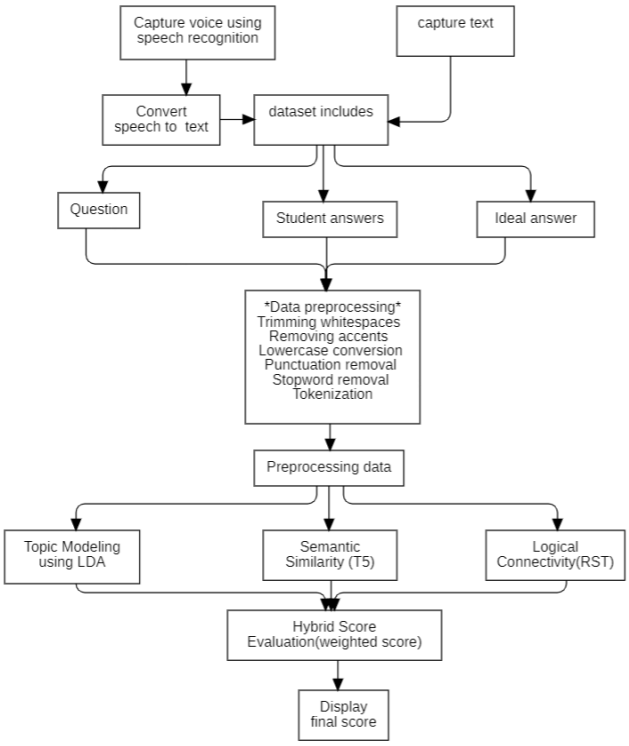

Fig: System architecture of our proposed work

IV. Results and Analysis

The voice-based paper evaluation system exhibits efficiency in assessing spoken responses through automation while still ensuring accuracy and fairness. Student answers were collected in Google Forms, on which the system was tested, and several performance metrics were utilized for evaluation. The hybrid scoring mechanism based on LDA for topic analysis, T5 transformer for semantic similarity assessment, and Rhetorical Structure Theory for coherence assessment worked salubriously to give meaningfully and meaningfully aligned score assignments. The degree of association between the assigned scores and the ones given by human scorers was found to be very strongly correlational from statistical evaluation using Pearson correlation and Mean Squared Error (MSE). The model was very effective in observing variations in student responses and logging assessment of logical structures.

With an evaluation of the confusion matrix, the system was confirmed as operating in accordance with relevance and structure to minimize errors in assessment. Evaluation using the BLEU score suggested the system measures linguistic accuracy, which compares the student response to the ideal answer, thus making evaluations objective. There was also qualitative input from faculty indicating that the system helped reduce grading time while ensuring consistency in scoring-so much so that biases that could find their way into the scoring of subjective answers could be eliminated.

Another important outcome demonstrates the system's adaptability to various speaking styles, accents, and minor differences in sentence structure while maintaining its accuracy. The interactive feedback mechanism further complemented students' educational processes by pointing to missing key points, logical flaws, and areas for improvement. The web-based user interface development using Gradio ensured a smooth experience for students and instructors in submitting, evaluating, and receiving feedback in real-time. The model further showed good generalizability across various types of questions, thus boosting the scalability of the solution to institutions.

The results, therefore, certify that the proposed system will prove beneficial as an implement operationally for the evaluation of subjective answers in an AI-guided format to lessen the burden of grading, enhance fairness, and restore ease in the lives of students.

Comparative Analysis

NO. Of Questions	Manual Evaluation(in marks)	Manual Accuracy(in %)	Automated Evaluation(in marks)	Automated Accuracy(in %)
3(5 marks each)	9	60	12	80
5(5 marks each)	18	72	22	88
2(10 marks each)	16	80	18	90
6(4 marks each)	19	79	22	91

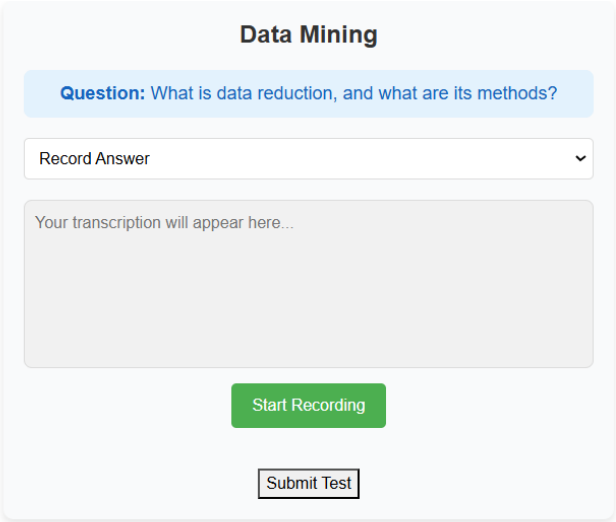

Fig: Recording the answer

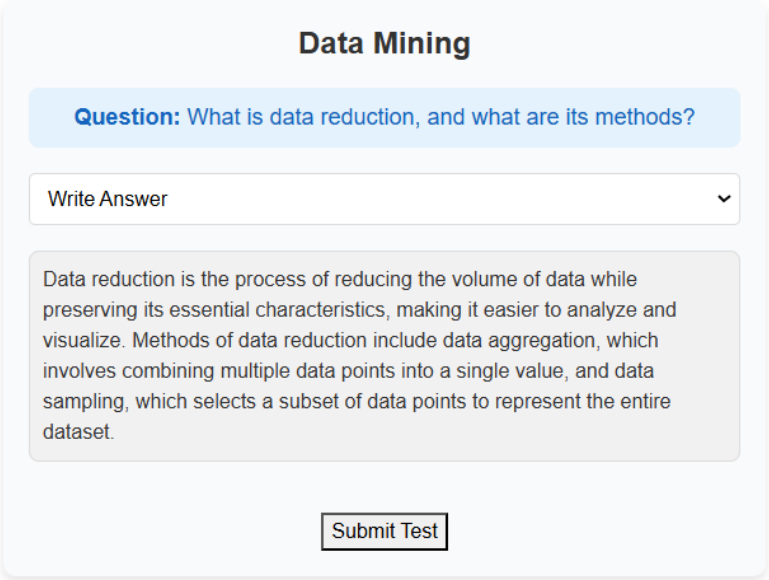

Fig: Conversion of recorded answer to text for evaluation

Test Results

Total Score: 8.67/20

Question	Student Answer	Score	Action
What is data reduction, and what are its methods?	data reduction up process of reducing the large data sets into the smaller one while having the relationships and patterns among themselves the main methods of data reduction are itch includes principal component analysis and compression techniques	7.67/10	Hide Reference
Reference Answer: Data reduction in data mining is the process of minimizing the size of large datasets while maintaining essential patterns and relationships. This helps improve the efficiency of data analysis. Common methods include dimensionality reduction, like Principal Component Analysis (PCA), which reduces the number of features without losing key information. Numerosity reduction techniques, such as clustering or sampling, compress data by representing it in a more compact form. Other methods include data compression, which reduces storage space, and aggregation, where data points are summarized to simplify analysis while preserving the overall trends. These techniques are essential for managing large-scale datasets and enhancing the performance of mining algorithms.			
Explain the concept of association rule?		1/10	Hide Reference
Reference Answer: Association rules in data mining are used to discover relationships between variables in large datasets. They consist of an antecedent(if part) and a consequent (then part), showing how the occurrence of one item or event is associated with the occurrence of another. These rules are evaluated based on metrics like support, confidence, and lift to identify strong and meaningful associations, often used in market basket analysis.			

Fig: Evaluating score and student answer vs reference answer

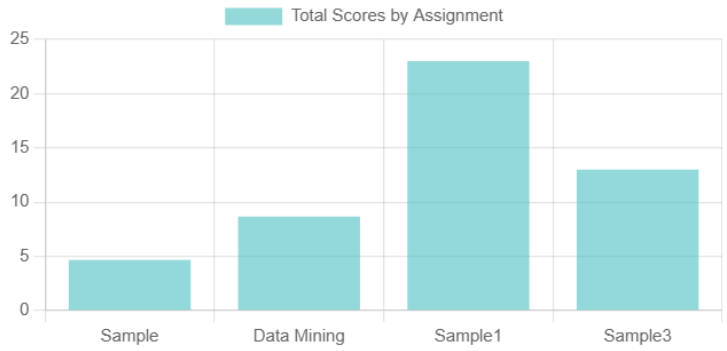

Fig: Progress chart of student over time

V. Conclusion

This project has ultimately built an automated voice-based evaluation of papers using the most advanced Natural Language Processing (NLP) technology along with deep learning. The combination of LDA along with T5 transformer and RST for the evaluation gives a diagnostic output on student responses wherein the topics relevance, semantic and logical coherence were being tested by the system. Thus, it seemed a really good system in place of the conventional manual grading system with faster, unbiased, and consistent evaluations.

The interactive feedback encompassed in the system equips students with constructive ideas regarding their answers to assist them in improving comprehension of the content. A very intuitive web interface submits answers, generates feedback, and therefore makes it accessible to the students and faculty alike.

This was an actual success story, but then there is still a lot of room for improvement. The inclusion of a comprehensive dataset from different fields in education, multilingual adaptation of the model, contextual understanding for more advanced responses, and some adaptive learning techniques would all make this system robust. Adaptive learning techniques could be included for generating student-specific feedback, increasing the personalized learning aspect of the system.

This is indeed an outstanding move toward AI-enabled instructional assessment. The project also indicates how technology can promote proper subjective evaluation through efficiency, accuracy, and fairness. And who knows, after a while, this perfectly improvised system could put an end to the age of assessment of spoken and written answers in educational institutions all over the world.

Future Scope:

The future vision for the project, however, lies in its expansion capabilities to include handwriting recognition features, programmatic assessments, and mathematical function evaluation, thus ensuring that a broader audience can be reached with complete diversity in education. Handwritten Answer Recognition can be accomplished using Optical Character Recognition (OCR) and deep learning models to transcribe handwritten information into the digital format, allowing students to write their answers as per their style of writing as useful in the subject that needs diagrams and equations. Programming-based evaluation might include the automated assessment systems for code, where the submitted code would be compiled, executed, and evaluated through predetermined test cases, logic analysis, and plagiarism to achieve reliable marking and instantaneous feedback in coding tasks. Mathematical function evaluation can integrate symbolic computation tools analyzing mathematics expressions, equations, and the like in a step-by-step approach rather than checking the ultimate answer, thus making it extremely relevant for the subjects of mathematics, physics, and engineering. Also, other eventualities can be included, such as graphical analysis tools, which would be used in geometry-based problems and graph plots. In that regard, they will be widening the possible flexibility of the system for different subjects open to technology and will also make digital assessments more accessible and fair.

References

[1] Md. Motiur Rahman, & Fazlul Hasan Siddiqui. (2018). A Natural Language Processing (NLP) based system for evaluating answer scripts.

[2] Sharma, Chahat, Bishnoi, Akash, Sachan, Akshay Kr, & & Verma, Aman. (2019). A novel automated essay evaluation system using NLP.

[3]Lim, Chun Then, & Bong, Chih How, Wong, Wee Sian, & Lee, Nung Kion. (2021). A systematic review of Automated Essay Scoring (AES) systems.

[4] Balaha, Hossam Magdy, & Saafan, Mahmoud M. (2021). AECF: AI-driven grading framework for MCQs, essays, and equation matching..

[5] Bashir, Muhammad Farrukh, Arshad, Hamza, Javed, Abdul Rehman, Kryvinska, & Natalia, & Band, Shahab S. (2021). Machine learning and NLP framework for subjective answer evaluation

[6] Wankhade, Mayur, Rao, & Annavarapu Chandra Sekhara, & Kulkarni, Chaitanya. (2022). Review of sentiment analysis methods and challenges.

[7] Kaseb, , Mostafa R, Rslan, Esraa,, Badry, & Rasha M., & Ali, Mostafa. (2023). Short answer question evaluation system for the Arabic language.

[8] Prabakaran, N., Kannadasan, R., Krishnamoorthy, & Kakani, & Vijay. (2023). Fully automated script evaluation system using Bi-LSTM and keyword-based algorithms.

[9] Tobler, Samuel. (2024). AI-driven assessment system incorporating generative AI models for response evaluation.

[10] Xu, Wenbo, Mahmud, Rohana, & Wai Lam. (2024). Feasibility and dependability of Automated Essay Scoring (AES) systems in education.

[11] Chandrapati, Lalitha Manasa, & Rao, Ch. Koteswara. (2024). NLP-based system for automatic assessment of descriptive answers.

[12] Grévisse, Christian. (2024). Large Language Models (LLMs) for grading short answer questions in medical education.

[13] Mohanraj, G, Nadesh, R. K.,, M., & Sathyapriya, V., & Marimuthu. (2024). NLP and deep learning framework for automatic answer script evaluation.

[14] Janda, Harneet Kaur, Pawar, Atish, Du, Shan,, & & Mago, Vijay. (2019). Comprehensive essay grading framework integrating syntax, semantics, and sentiment analysis.

[15] Pimpalshende, Anjusha, and A. R. Mahajan. "Test model for stop word removal of Devanagari text documents based on finite automata." 2017 IEEE International Conference on Power, Control, Signals and Instrumentation

Engineering (ICPCSI). IEEE, 2017.2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI). IEEE, 2017.

[16] Pimpalshende, A. N. "Overview of text summarization extractive techniques." International Journal of Engineering and Computer Science 2.4 (2013): 1205-1214.

[17] Pimpalshende, A., and A. R. Mahajan. "Extraction of root words using morphological analyzer for Hindi text." Int J Soft Comput 13.5 (2018): 134-138.

[18] Pimpalshende, Anjusha, and A. R. Mahajan. "Pre-processing phase of Hindi language text summarization System." International Journal of Computer Science and Information Security (IJCSIS) 14.5 (2016).

[19] Potnurwar, A., Pimpalshende, etl (2020). Extractive multi-document text summarization by using binary particle swarm optimization. Helix, 10(04), 263-265.

[20] Binary Particle Swarm Optimization with an improved genetic algorithm to solve multi-document text summarization problem of Hindi documents SS Aote, A Pimpalshende, A Potnurwar, S Lohi - Engineering Applications of Artificial Intelligence, 2023.