IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 30 April 2024, accepted 16 June 2024, date of publication 21 June 2024, date of current version 28 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3417706

RESEARCH ARTICLE

Descriptive Answers Evaluation Using Natural Language Processing Approaches

LALITHA MANASA CHANDRAPATI AND CH. KOTESWARA RAO, (Member, IEEE)
School of Computer Science and Engineering, VIT-AP University, Amaravati 522237, India
Corresponding author: Ch. Koteswara Rao (koteswararao.ch@vitap.ac.in)

ABSTRACT Answer scripts are an important aspect in evaluating student's performance. Evaluating papers from a descriptive outlook can be a challenging and exhausting task. Typically, answer script evaluations are conducted dynamically, which can lead to bias and can be quite time-consuming. Various efforts have been made to automate the evaluation of student responses with the usage of Artificial Intelligence techniques. Yet, most of the work relies on particular words or typical counts to accomplish this task. In addition, there is a shortage of organized data sets too. In this research a novel ensemble model Descriptive answer evaluation system(DAES) is introduced, which integrates Topic Modelling (TM) and Question Answering (QA) models for automatically evaluating descriptive answers. Latent Dirichlet Allocation (LDA) and a fine-tuned Text-to-Text Transfer Transformer(T5) models were utilized to identify key topics and the correctness of specific statements within the student answers. Sentence-BERT is utilized to encode sentences and cosine similarity method is applied to generate similarity scores. For this approach, LDA studies thematic evaluation, T5 assess for semantic analysis of the student answer. A final score is given to each answer after a thorough review procedure using predetermined criteria. Experiments results in achieving an accuracy of 95%, precision of 94%, recall 95% and f1-score of 94% on training data by using the proposed model.

INDEX TERMS Descriptive answer evaluation, LDA, natural language processing, T5, sentence similarity.

I. INTRODUCTION

Evaluating a student's performance and abilities through descriptive questions and answers provides an in-depth assessment. These answers allow students to freely voice their concepts and comprehending of the subject matter without any restrictions. Nevertheless, there are noticeable variations between descriptive and objective answers. Descriptive answers often require more time to formulate due to their lengthier nature. In addition, they involve additional contextual information, which demands increased focus and neutrality from the evaluator.

Assessing such kinds of questions using Artificial Intelligence poses a difficulty because of the underlying ambiguity in natural language. Several preprocessing steps are required prior to analysis, such as data cleaning and tokenization. Various techniques can be used to compare textual data, including methods like latent semantic structures, ontologies, document similarity and concept graphs. In order to determine the final score, various factors need to be taken into account. These factors include similarity, the presence of specific elements, and language considerations [1], [2]. Previous efforts have been made to address this issue, but there is still potential for enhancements [3], [4], [5], which will be explored in this research.

Descriptive examinations are often perceived as increasingly challenging and intimidating by both teachers and students because of their inherent characteristic: context. The accuracy of an answer relies heavily on the thoroughness of the evaluator, who must carefully evaluate each word for scoring. The evaluator's mental well-being, level of fatigue, and ability to remain objective greatly impact the outcome. Thus, it is far more efficient to delegate the descriptive answers evaluation system, saving time and resources. Assessing objective answers using machines is a simple and possible process. To optimise the analysis of student's responses, a procedure may incorporate brief responses to queries. However, addressing descriptive answers can

The associate editor coordinating the review of this manuscript and approving it for publication was Wei Ni.

be quite difficult. The lengths of these texts vary and they encompass a wide range of vocabulary. In addition, individuals tend to utilise alternative words with easy abbreviations, which adds complexity to the process.

A significant amount of effort has been focused on evaluating subjective answers using different methods. Text analysis can be conducted using various methodologies. These encompass evaluating the similarity between words and texts, deriving potential solutions from the contextual meaning of the text, quantifying noun phrases in documents, and identifying recurring keywords in responses. However, there are still ongoing challenges that need to be addressed. These include the problem of losing semantic context in term frequency-inverse document frequency (TF-IDF) [6], the need for better tuning of hyperparameters [7], the resource-intensive nature of training [8], and the requirement for improved datasets [5].

In this work, an approach that utilises natural language processing(NLP) and deep learning(DL) techniques to evaluate descriptive answers was implemented. Our work employs NLP techniques, including Tokenization, Lemmatization, word embedding techniques as TF-IDF, similarity methods as Cosine Similarity, LDA model for topic modelling, T5 for question answering model. Multiple metrics for evaluation were used such as Accuracy, Precision, Recall, F1-Score, Similarity scores. Other approaches were also considered that have previously been used to evaluate descriptive answers or estimate text similarity in general.

TM technique, LDA allow examiners to uncover the underlying themes and concepts present in student's written responses. By analysing the topics extracted from student answers, educators can gain insights into student's understanding, strength and weakness. QA models, powered by state-of-the-art deep learning architectures such as T5, enable automated evaluation of student's ability to comprehend and respond to questions. Integrating topic modeling and question answering enables a comprehensive assessment, considering both the accuracy of student answers and their coherence and relevance to underlying topics.

This research proposes a novel ensemble model DAES to evaluate student descriptive answers efficiently utilizing deep learning and NLP approaches. The model emphasizes the evaluation resulting in robust outcome of the student answer scripts. First, the LDA model is applied to both Ideal Answer, Student Answer. The topics obtained respectively are compared for similarity score. The second step involves deploying of T5 model on both Ideal answer, Student answer and the predicted answers given by model are undergone similarity check. Further, both LDA score and T5 score are aggregated to give the final score.

Let consider an example,

Question:"What is the purpose of a binary search algorithm?"

Ideal Answer: "By partitioning the search range in half iteratively, a binary search algorithm attempts to locate desired value in an efficient manner within a sorted array."

Student Answer: "The purpose of a binary search algorithm is to find the position of a target value by repeatedly dividing the search interval in half within a sorted array."

Using LDA, evaluators can analyse the responses of the Ideal answer, Student Answer to identify underlying topics. The model might identify topics related to algorithms, efficiency, and searching for the ideal answer and Student Answer, indicating similarity in understanding. Employing T5, examiners can assess the accuracy and depth of student's responses. When prompted with the question, the model would correctly identify the ideal answer as accurately describing the purpose of binary search. Similarly, Student's response would also be identified as accurate. LDA scores, T5 scores are calculated and combined to give final score.

A. MOTIVATION

This method helps in providing educators and institutes quality assessment approach which they can rely on, reduces the time for evaluation and provide effective results. Traditional evaluation methods often relies on factual knowledge or problem-solving skills. Integrating topic modeling and question answering offers a more holistic approach, considering both the accuracy of responses and their alignment with underlying concepts. Hence, teachers can spend more time in teaching, creating curriculum and other activities which enhances better education for students.

B. CONTRIBUTIONS

The notable contributions of our research reside in its novel methodology and the progress it introduces to the domain of automated assessment of student responses.

- Implementing a novel hybrid model that integrates TM model (**LDA**) and QA model (**T5**) techniques for evaluating student answers.
- Integrating thematic coverage and semantic understanding in a single framework, allowing for a more comprehensive assessment of student responses.
- Enhancing efficiency and scalability in the evaluation process by automating the assessment of student answers, decreasing the amount of time and effort spent on manual evaluation.
- Providing consistent and objective evaluation of student answers by minimizing subjective biases and inconsistencies inherent in human grading.

C. PAPER ORGANIZATION

The subsequent sections of the paper are structured in the following manner: Section II provides a literature review, Section III outlines the methodology proposed, Section IV includes the experimental setup and discusses the results, and Section V concludes the paper.

II. LITERATURE REVIEW

A. TRADITIONAL EVALUATION METHODS

Traditional assessment methods in education include a variety of methodologies for measuring student comprehen-

sion, knowledge, and abilities within a certain subject or curriculum. Multiple choice questions, short answer questions, essays, fill-in-the-blank exercises, true/false questions, matching activities, oral examinations, practical tests, and peer evaluations are all common assessment methods. Each of these methodologies assesses distinct areas of student learning, ranging from basic understanding to critical thinking, problem solving, and practical application. Manual grading requires teachers to thoroughly analyse and assign grades to student work based on predefined criteria, assuring fairness and consistency. Rubrics give a systematic framework for evaluation by dividing assessment criteria into distinct components and degrees of performance.

Several studies have used the strategy, Statistical technique which depends on keyword matching and is limited in its ability to account for synonyms and context for descriptive paper assessing [9], [10]. Information extraction methods deconstruct text into concepts and relationships by recognising a structure or pattern in the text [11]. Dependencies create a substantial impact on the formulation of scores and should be verified by domain experts [12], [13].

B. AUTOMATED EVALUATION APPROACHES

In response to the difficulties encountered by human evaluation systems while evaluating descriptive student responses, the authors [14] suggested a unique automatic assessment system based on a syntactical relation-based feature extraction approach. Furthermore, the system has implemented a cognitive-based methodology, whereby the accuracy of student responses is assessed according to the phrases employed to respond to the queries.The overall rating and accompanying comments serve as an indicator of the subject's level of knowledge. In comparison to existing grading systems, the implementation outperforms them by 95% accuracy, 94% recall, and 94.5% sensitivity. Furthermore, combining grammar analysis and fingerprints may improve accuracy by evaluating the language details and context provided in student answers. Limitations might include the necessity for ongoing refining to adapt to changing language usage and the difficulty of effectively capturing semantic details.

In their study, the authors [15] explains an experiment in which handwritten descriptive responses on Japanese language university entrance examinations were scored entirely automatically. The proposed methodology integrates handwriting recognizers based on deep neural networks that were trained on labelled data with a language model constructed from an extensive generic corpus. Character accuracy surpasses 97%, and the Quadratic Weighted Kappa (QWK) score varies from 0.84 to 0.98, showing a strong match to human examiner assessment. The pipeline uses advanced deep neural networks for character recognition and automated scoring, delivering great accuracy even with few labelled patterns. Training data from the ETL database, together with multiple transformations and fine-tuning on the NCUEE-HJA1K dataset, guarantees reliable performance. The study emphasises the effectiveness and dependability of automatic scoring methods, addressing the time-consuming aspect of analysing handwritten responses. Furthermore, while the SCUT-EPT dataset presents hurdles for text recognition algorithms, human examiners now outperform sophisticated text recognition techniques such as CRNN and attention mechanisms on this dataset. These findings highlight the need for more study on end-to-end automatic scoring of descriptive replies, as well as the need of improving recognition and scoring procedures for handwritten evaluations.

The authors [16] describes Topic-aware BERT, a unique approach to automated essay scoring (AES) and key topical sentence (KTS) retrieval. This method uses self-attention maps in intermediary layers to determine the link between essay scores, student essays, and thematic information from essay directions. By using prompt-specific knowledge, Topic-aware BERT outperforms earlier neural-based AES techniques. It also efficiently recognises essential topical phrases in argumentative essays, which improves understanding of the essay material. The evaluation of Topic-aware BERT utilising open and manually annotated datasets indicates its competitive AES performance as well as the usefulness of the KTS retrieval approach. Notably, this study closes the gap between neural-based AES and automatic writing evaluation (AWE) systems, making an important addition to the area. Overall, topic-aware BERT is an option for strengthening AES and expanding the capabilities of AWE systems by incorporating topical information.

A new method for automatically assessing descriptive answers using multiple ML and NLP approaches was proposed by the authors [17]. The study uses Word2vec, multinomial naive Bayes (MNB), cosine similarity, Word Mover's Distance (WMD), WordNet, and TF-IDF to assess answers based on answer statements and keywords. A machine learning model is built to predict answer grades, and WMD outperforms cosine similarity. Without MNB, the model achieves 88% accuracy, but MNB decreases the error rate by 1.3%.

Wagh and Anand [18] suggested a multi-criteria decision-making approach to assess the similarity of legal documents. The study utilized AI and aggregation approaches like ordered weighted average (OWA) to determine the similarity of papers. The dataset includes Indigenous People Supreme Court case judgements from 1950 to 1993. Recall along with F1score were utilised as evaluation metrics. The suggested concept-based similarity method outperformed previous strategies, including TF-IDF, with F1-scores of up to 0.8.

Alian and Awajan [19] determined sentence context by analysing paraphrasing and phrase similarity with clustering algorithms, weighting methods and word embedding models. FastTex and AraVec are Arabic embeddings that have been pre-trained. The Arabic dataset comprised approximately

77,600,000 tweets. Pre-trained embedding using labelled data from experts resulted in higher recall and accuracy for K-means and agglomerative clustering (0.87% and 0.78%, respectively). A concept graph-based technique was suggested by Jain and Lobiyal [20] to evaluate subjective questions. For both the solution and the response, concept graphs were produced, and the score determined utilising methods graph similarity methods. Montes-y-Gómez et al. [21] described approaches for identifying similarities across idea graphs and retrieving information from them.

Bahel and Thomas [22] proposed a framework for evaluating descriptive questions using text semantics, keyword summarization, text summarizing and comparing the findings to current methods. The findings indicated an inaccuracy of 1.372, compared to 1.312 for Jaccard's similarity technique. The technique did not work for nontextual data like diagrams, pictures, and other forms. Zhang and Litman [23] attempted to obtain AES feedback by analysing intermediate layers of deep neural networks. Zhang and Litman used LSTM-based coattention neural networks to identify topical components in a writing assignment called response-to-text assessment (RTA).

The model proposed by Darwish and Mohamed [39] utilises Latent Semantic Analysis (LSA) and fuzzy ontology to improve the assessment of essays by taking into account both semantic content and coherence. This approach offers a reliable and unbiased alternative to manual grading. The main findings emphasise the system's effectiveness in evaluating essays and providing valuable feedback. However, there are limitations such as potential difficulties in handling extremely complex language and the requirement for customisation specific to domains.

In the research Süzen et.al. [40] explores the application of data mining techniques and clustering to quantify the similarity between student responses and a reference answer. The objective is to facilitate automated grading and feedback. The major findings emphasise the capacity of computational methods to assist and improve human rating. Nevertheless, this approach has certain constraints such as its dependence on frequently used phrases, the possibility of oversimplification in categorization, the requirement for human adaptation, and its restricted applicability.

Table 1 gives the summary of the prior work done. There are several inherent limitations to the wide range of automated assessment approaches that are indicated in the Table 1. Although strategies like syntactical relation-based feature extraction and cognitive-based assessment have opportunities to improve grading accuracy, there are still difficulties in modifying these methods to account for changing language usage and capturing minute semantic details. Furthermore, scalability and generalisation issues may arise from depending on deep neural networks and language models for automated scoring, as demonstrated in studies using Topic-aware BERT and deep neural network-based handwriting recognizers. In addition, the application of ML and NLP techniques–such as WordNet, Word2vec, and Multinomial Naive Bayes–may encounter challenges when processing complicated language structures and require domain-specific customisation to achieve optimal performance. Despite progress, difficulties still arise when managing huge data sets, interpreting results, and resolving the drawbacks of existing assessment approaches. These issues highlight the necessity for ongoing study and improvement of automated assessment methods.

C. GAP ANALYSIS

The evaluated literature provides a complete overview of both conventional and automated assessment methods in education, emphasising the variety of methodologies used to measure student learning. Traditional evaluation techniques include a variety of formats, ranging from multiple choice questions to essays and oral examinations, each designed to examine distinct areas of student knowledge and skills. Automated evaluation systems, on the other hand, using DL and NLP techniques to speed the assessment process while also providing scalability and consistency. While these automated methods show promise in terms of efficiency and accuracy, there are significant gaps in study. These include the need to investigate innovative assessment methods beyond traditional and automated techniques, specificity in evaluating different academic disciplines, integration of human expertise with automated systems, inclusion of multimodal data, and scalability and generalisation of evaluation models across diverse contexts. It is of utmost importance to address these existing gaps in order to propel the field of educational assessment forward and establish comprehensive and fair evaluation practices in the coming years.

III. PROPOSED METHODOLOGY

The proposed approach comprises a data collecting and annotation module, a preprocessing module, a topic modelling model, a question answering model, a sentence embedding module, a similarity measurement module, and an evaluation and scoring module. Initially, the user provides inputs, including questions, student answers, and ideal answers. Figure 1 depicts the framework proposed.

A. IDEAL ANSWER

An ideal answer is a descriptive answer that is utilised to categorise student's answers. The response should encompass all the keywords and contextual information provided, organised in separate paragraphs or lines or for clarity and coherence. The teacher/evaluator typically generates the ideal answer to the given question.

B. STUDENT ANSWER

It is the descriptive answer provided by the learner and requires to be assessed. Typically, it includes a selection of the keywords and ranges from one or several sentences, based on question nature and student style of writing. The text typically includes synonyms more frequently than the ideal answer, thus necessitating greater attention to semantic processing.

TABLE 1. Summary of literature.

Ref. No.	Methodology Used	Key Findings	Limitations
14	Syntactical relation-based feature extraction; Cognitive-based approach	Outperforms existing grading systems with 95% accuracy, 94% recall, and 94.5% sensitivity; Combining grammar analysis and fingerprints may improve accuracy; Addressing changing language usage and capturing semantic details required	Necessity for ongoing refining; Difficulty in effectively capturing semantic details
15	Deep neural network-based handwriting recognizers; Language model from generic corpus	Character accuracy exceeds 97%; Quadratic Weighted Kappa (QWK) score ranges from 0.84 to 0.98; Effective in automating scoring of handwritten responses; Emphasizes the effectiveness and dependability of automatic scoring methods	Challenges in text recognition algorithms; Need for more study on end-to-end automatic scoring of descriptive replies; Improving recognition and scoring procedures for handwritten evaluations
16	Topic-aware BERT	Outperforms earlier neural-based AES techniques; Efficiently recognizes essential topical phrases in argumentative essays; Competitive AES performance; Closes the gap between neural-based AES and AWE systems	Potential scalability and generalization issues; Further research needed on integration into existing educational systems
17	MNB, TF-IDF, Word Mover's Distance, Word2vec, Cosine similarity, WordNet	WMD outperforms cosine similarity; Achieves 88% accuracy without MNB; MNB reduces error rate by 1.3%; Utilizes multiple ML and NLP approaches for answer assessment	Potential limitations in handling complex language structures; Need for domain-specific customization; Scalability concerns
18	Multi-criteria decision-making approach	Concept-based similarity method outperforms previous strategies; Utilizes AI and aggregation techniques like OWA; Achieves F1-scores of up to 0.8	Challenges in handling large datasets; Interpretation of results may require domain expertise
19	Word embedding models, clustering algorithms, weighting methods	Pre-trained embeddings improve recall and accuracy; Concept graph-based technique suggested for evaluating subjective questions; Describes approaches for identifying similarities across idea graphs	Potential limitations in handling diverse data formats; Need for further refinement and validation
20	Concept graph-based technique	Provides a method for evaluating subjective questions using concept graphs; Scores determined using graph similarity approaches	Potential challenges in scaling up to large datasets; Interpretation of graph-based similarity measures
21	Approaches for identifying similarities across idea graphs	Describes techniques for identifying similarities across idea graphs and retrieving information from them	Potential challenges in handling complex graph structures; Need for efficient algorithms for graph comparison and retrieval
22	Text summarizing, text semantics, keyword summarization	Descriptive questions evaluation framework proposed; Comparative analysis with Jaccard's similarity technique; Finds inaccuracy compared to Jaccard's method; Limitations in handling non-textual data formats	Difficulty in handling non-textual data; Need for enhanced accuracy in evaluation methods
23	Analysis of intermediate layers of deep neural networks	LSTM-based coattention neural networks for AES feedback; Identifies topical components in writing assignments; Focus on response-to-text assessment (RTA)	Potential scalability and generalization issues; Need for integration into existing educational systems; Interpretation challenges

C. DATA COLLECTION AND ANNOTATION

In order to train and evaluate the proposed framework, a substantial corpus consisting of descriptive question answers is required. However, As per our knowledge it is understood that there is currently no publicly accessible annotated corpus of descriptive question answers. In this study, a corpus was developed which consists of descriptive answers that have been annotated. In order to generate this corpus, web crawling was conducted on multiple websites to gather a corpus of descriptive question responses. The crawled data encompasses computer science.

The collected data is unlabelled and requires annotation. To achieve this objective, a team of volunteers was assembled who possess expertise in the same field as our corpus. 20 annotators were engaged who are our acquaintances and colleagues from various colleges in India, with most of them being teachers. Annotators were assigned the responsibility of evaluating the answers provided by students and determining the most accurate score.

1) DATASET STATISTICS

Our dataset contains 300 questions, their corresponding 300 ideal answers and 30 student's answers for each question. All the answers are short descriptive answers. All the questions belong to Computer science domain. The questions and answers used in the dataset were representative of typical exam scenarios. The lengths of the questions and responses were consistent with what one would expect in real-world academic settings. This consideration ensures that the model's evaluation process is applicable to actual educational contexts. while the initial study involved a small sample size, the results are valid as a proof of concept. The use of pre-trained models, cross-validation techniques, and careful

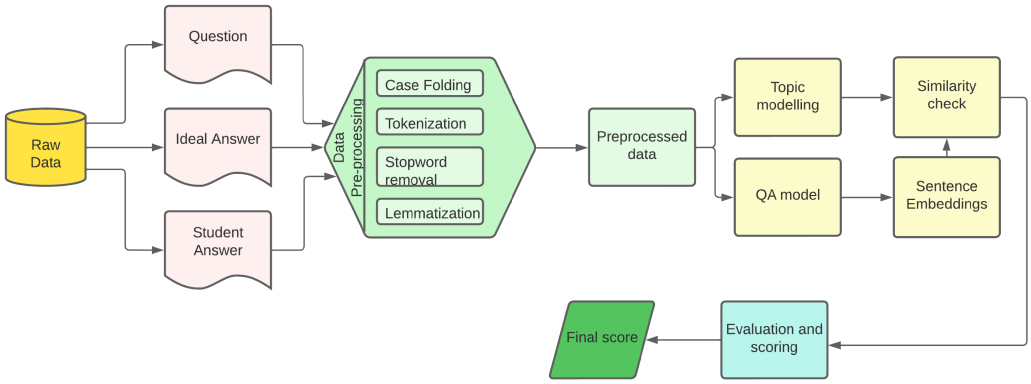

FIGURE 1. DAES proposed system.

consideration of question and answer lengths provide a solid foundation for the proposed solution.

The size of the question pool was carefully chosen to provide broad coverage and instructional relevance while being manageable for extensive examination. The actual dataset contained 30 student responses per question, giving a solid foundation for dependable and significant results. The Table 2 depicts the dataset sample, where 5 questions, their corresponding ideal answers and 2 student answers for each of the 5 questions were given for illustrative purpose only. Like wise there are 300 questions and their corresponding set of ideal and student answers.

D. DATA PREPROCESSING

Text documents need to go through preprocessing to prepare them for machine processing. This stage is known as preprocessing which involves employing several NLP techniques, including Case Folding, Tokenization, Lemmatization, Stopwords Removal, temming, and Parts of Speech(POS) Tagging. An overview is provided for some of these strategies.

1) TOKENIZATION

The most crucial and first step in the process of NLP is Tokenization [27]. It divides text into tiny components, like characters, words, sentences and paragraphs. It involves the separation of each word to accurately determine its intended significance. In this study, the data is segmented into sentences and words by utilising period marks, spaces. Sirts and Peekman [28] present an analysis of the evaluation outcomes for three established methods for word tokenization, phrase segmentation for the Estonian web dataset.

2) STOPWORDS REMOVAL

Natural Language encompasses a wide range of words, including essential features like ‘the’, ‘in’, ‘on’, ‘is’, and others, which facilitate human comprehension. These phrases are typically inconsequential in the majority of deep learning applications and have the potential to impede the training process by supplying superfluous data to the model. Each language often contains a collection of frequently used stop words that are commonly eliminated from the corpus in order to enhance the density and distinctiveness of the dataset. Schofield et al. [29] contend that the removal of stopwords is merely superficial and that eliminating stopwords has a negligible impact on improving topic inference, save for common phrases. Çagatayli and Çelebi [30] discovered that removing stopwords has a negligible impact on the actual outcomes. Nevertheless, it is essential to acknowledge that removing frequently occurring terms that have minimal significance can enhance the effectiveness of the machine learning model.

3) PARTS OF SPEECH TAGGING

POS tagging is the process of assigning a specific part of speech, such as a noun, verb, adverb, or adjective, to each word in the given data. Various tools can be utilised for part-of-speech annotation, such as the NLTK POS tagger, which aids in comprehending the sentence’s structure. It can be used to identify noun phrases inside a sentence, reduce terms to their lemma, and for various other fascinating uses. Divyapushpalakshmi and Ramalakshmi [31] utilised POS tagging to enhance the effectiveness of sentiment analysis on the Twitter network.

4) LEMMATIZATION

The process of transforming words in a dataset into it’s base or root form is called Lemmatization. It’s particularly useful for capturing variations like different tenses. For instance, ‘write’, ‘writing’, and ‘written’ all stem from the same root ‘write’. To perform lemmatization effectively, a comprehensive dictionary is necessary to map terms to their corresponding base forms, which are known as lemmas. This process utilizes information about the part of speech of each word to ensure accurate mapping. In a study by Camastra and Razi [32], they utilized Lemmatization in combination with support vector machines to classify Italian texts.

5) STEMMING

Stemming is a method aimed at reducing words to their core stems, operating on the principle that languages

TABLE 2. Sample list of a dataset.

Qid	Question	Sid	Student Answer	Ideal Answer
1	Explain the concept of object-oriented programming (OOP) and provide an example.	s1	Object-oriented programming (OOP) is a programming paradigm that focuses on organizing code into objects, which are instances of classes. These objects can encapsulate data and behavior, making it easier to model real-world entities. For example, a "Car" class can have attributes like "make," "model," and "year," along with methods like "start engine" and "accelerate."	Object-oriented programming (OOP) is a programming paradigm that focuses on organizing code into objects, which are instances of classes. These objects can encapsulate data and behavior, making it easier to model real-world entities. For example, a "Car" class can have attributes like "make," "model," and "year," along with methods like "start engine" and "accelerate."
		s2	OOP stands for Out-of-Place programming, which means writing code in a location different from where it's intended to be used.	
2	Describe the difference between a compiler and an interpreter.	s1	A compiler is a type of software used to browse the internet, while an interpreter is a device used to control hardware components.	A compiler and an interpreter are both language translators, but they operate differently. A compiler translates the entire source code into machine code before execution, producing an executable file. An interpreter, on the other hand, translates the code line by line at runtime. This means an interpreter executes the program directly, without generating an independent executable.
		s2	A compiler translates the entire source code into machine code before execution, resulting in an independent executable file. An interpreter, on the other hand, translates the code line by line at runtime.	
3	What is the purpose of a database index, and how does it improve query performance?	s1	A database index is a tool used for creating visual representations of data.	A database index is a data structure that improves the speed of data retrieval operations on a database table at the cost of additional storage space and decreased performance on data modification operations. It works by creating a sorted list of key values, allowing the database engine to quickly locate the rows that satisfy a given condition. This leads to faster query performance, especially on large datasets.
		s2	A database index is a data structure that improves the speed of data retrieval operations on a database table by creating a sorted list of key values. This allows the database engine to quickly locate the rows that satisfy a given condition, leading to faster query performance.	
4	Explain the concept of Big O notation and its significance in algorithm analysis.	s1	Big O notation is a mathematical notation used to analyze the efficiency of algorithms. It helps us understand how the time or space requirements of an algorithm scale with the size of the input.	Big O notation is a mathematical notation that describes the limiting behavior of a function when the argument tends towards a particular value or infinity. In computer science, it is used to analyze the efficiency and complexity of algorithms. For example, O(n) represents linear time complexity, indicating that the time taken by the algorithm is directly proportional to the size of the input.
		s2	Big O notation is a mathematical notation used in computer science to describe the efficiency of an algorithm in terms of time or space complexity. It provides an upper bound on the worst-case performance of an algorithm. For example, O(n) represents linear time complexity, indicating that the time taken by the algorithm is directly proportional to the size of the input.	
5	Describe the process of memory allocation and deallocation in a programming language.	s1	Memory allocation is a method used for designing user interfaces in software development.	Memory allocation is the process of reserving a portion of a computer's memory for a specific purpose. This allows programs to dynamically allocate and deallocate memory during runtime. Memory deallocation, on the other hand, involves releasing the memory that was previously allocated, preventing memory leaks and efficiently utilizing system resources.
		s2	Memory allocation in programming involves reserving a block of memory for storing data. This can be done using functions like malloc() or new. Deallocation, on the other hand, involves releasing that memory back to the system using free() or delete.	

adhere to formal grammatical rules, generating vocabulary accordingly. By removing suffixes that differentiate related words, stemming can effectively simplify them. This process encompasses actions like transforming plurals into singular forms and trimming concluding characters. Across languages, diverse stemming techniques are available, with examples including Potter's algorithm designed for stemming English words. Jabbar et al. [33] explore a range of stemming approaches applicable to textual data.

6) CASE FOLDING

Natural language comprises words that often exist in various cases, with instances where the same word is repeated in accordance with its case. As a result, it is usual to convert all the data to the same case, typically lowercase, in order for the machine to understand each word consistently.

Once the data has undergone preprocessing in accordance with the specified criteria, textual data is transformed into a numerical format. This is necessary since machines are only capable of comprehending numbers and do so with great proficiency. The practice of representing words as numerical vectors is known as word embedding. Some of the approaches utilised in this process include Bag of Words (BoW), TF-IDF, and word2vec. TF-IDF was employed for the purpose of word embedding in our work.

7) TF-IDF

TF-IDF is similar to BoW in that it tallies the occurrences of all words within a text. However, it goes a step further by factoring in the number of unique sentences that contain those words. Consequently, it offers insights into both the frequency and the importance of a word within the document. Sammut and Webb [34] conduct an extensive examination of TF-IDF, while Havrlant and Kreinovich [35] offer a probabilistic interpretation of this technique. Additionally, Thakkar and Chaudhari [36] utilize TF-IDF for forecasting stock trends.

E. PROPOSED FRAMEWORK

To evaluate students' descriptive answers effectively, topic modeling and question answering systems are integrated. LDA (Latent Dirichlet Allocation) is employed to assess thematic coverage, and a finetuned T5 model is used to evaluate the semantic understanding of the student answers. The detailed process is explained in Algorithm 1.

Topic modelling, specifically employing methods such as LDA, aids in identifying the fundamental themes or topics that exist within a collection of textual data. Through the examination of word distribution in documents, LDA can detect prominent themes or topics that are present in both the student's answer and the ideal answer. The degree of similarity between the topic distributions of the student's answer and the ideal answer serves as an indicator of how effectively the student has addressed the desired thematic features in their answer. For instance, if the ideal response highlights specific essential concepts or themes, a strong student answer should also address these themes, and LDA can measure the thematic similarity between the student and ideal answers.

Process:

- Apply LDA to both the student answers and the ideal answers to extract thematic topics.
- Represent each answer as a distribution over the identified topics, Student answer topic distribution $S_{\text{topic\_dist}}$. and Ideal answer topic distribution $I_{\text{topic\_dist}}$.
- Compute the similarity cos_sim_{lda} using cosine similarity, between the student answer topic distributions and ideal answer to quantify thematic similarity.

Question answering models, like T5, are specifically engineered to comprehend the contextual correlation between questions and answers. T5 has the ability to produce answers to queries by taking into account the context given in the question and providing appropriate solutions. T5's understanding of semantics is primarily oriented towards generating coherent and contextually appropriate responses given input text. While T5 inherently captures semantic aspects during fine-tuning on tasks like question answering, its primary focus is on generating text rather than assessing semantic similarity between existing text pairs. To overcome this Sentence embedding model Sentence-Bert(SBert) was utilized to extract embeddings for both T5 generated student answers and ideal answers and then compare them for similarity evaluation.

FINETUNING T5: For our model, ''t5-small'' variant(60 M parameters) was considered. The model was pretrained on the Stanford Question Answering Dataset (SQuAD) [41]. During pre-training, the T5 model is optimized to generate accurate answers given context paragraphs, with hyperparameters carefully selected for optimal performance. With a learning rate of 3e-5, batch size of 32, AdamW optimizer the model is evaluated. Following pre-training, the model transitions to fine-tuning on our dataset containing student answers, ideal answers and questions, aiming to adapt its capabilities. Hyperparameters for fine-tuning, including a learning rate of 2e-5, batch size of 16, are adjusted to suit the characteristics of the dataset.

Process:

- Fine-tune the T5 model on the exam questions, student answers, and ideal answers to capture semantic understanding.
- Use the fine-tuned T5 model to generate answers for both the student and ideal questions.
- Generate embeddings for both t5 predicted student answers and ideal answers using Sbert model
- Compute the cosine similarity between the embeddings of the student's response E_{student} and the ideal response E_{ideal} to quantify semantic similarity.

The combined use of T5 and SBERT enhances the accuracy of the evaluation process by considering both the quality of the generated response and its semantic similarity to the ideal answer. This approach ensures a more comprehensive assessment, covering both the relevance and semantic fidelity of the student's response compared to the ideal answer.

1) TOPIC MODELLING

The LDA model, which was first introduced by Blei et al. [24], is a mathematical structure that aims to depict documents and topics as multinomial distributions over vocabulary and topics, respectively. LDA is able to identify topics within vast corpora through the examination of the co-occurrence frequencies of various terms. The process

involves simplifying each document into a word vector using the ''bag-of-words'' method. Subsequently, a term vocabulary is generated through the analysis of term frequencies within the document. By means of this procedure, the model discerns themes via an analysis of the probability distribution of words and allocates documents to topics in accordance with their probability.

The basic procedure of LDA is:

Generate document-topic distribution θ from Dirichlet prior parameterized by

Notations:

- D: number of documents
- N: number of words in each document
- K: number of topics
- V: size of the vocabulary (total number of unique words)

Generative Process:

- For each document d in D:
  - Choose a distribution over topics $\theta_d \sim$ Dirichlet(α).
  - For each word n in N:
    - Choose a topic $z_{d,n} \sim$ Multinomial(θ_d).
    - Choose a word $w_{d,n}$ from the topic $z_{d,n} \sim$ Multinomial($\beta_{z_{d,n}}$).

Parameters:

- α: Hyperparameter controlling the document-topic density. A higher value of α means documents are likely to be made up of more topics.
- β: Hyperparameter controlling the topic-word density. A higher value of β means topics are likely to contain a mixture of most words.

Topic models are generally estimated using two approaches for inferring parameters: variational Bayesian inference [24] and Gibbs sampling [25]. Gibbs sampling is a probabilistic procedure used to obtain samples from a Markov chain. Collapsed Gibbs sampling is a frequently employed method for estimating parameters in LDA. The benefit of LDA stems from its multi-assignment approach, allowing documents to be assigned to many subjects.

2) QUESTION ANSWERING APPROACH

T5, developed by Google Research, is an advanced natural language processing (NLP) paradigm also referred to as Text-to-Text Transfer Transformer (26). T5 is a transfer model that can be modified to perform question-answering, language translation, text classification, and other natural language comprehension tasks. By undergoing training on a vast collection of textual data, the model gains the ability to understand and generate a wide range of natural language.

T5 introduces a significant advancement in transfer learning through its utilization of a ''prefix'' methodology. This involves fine-tuning the model for a particular job by training it with an extra prefix to the input text. To fine-tune T5 for a text classification task, the input text should be prefixed with the task name and a separator, such as ''classify: This is the input text.''

This model utilizes a transformer-based architecture, with the encoder and decoder composed of many layers of self-attention and feedforward neural networks. T5 employs subword tokenization to segment input text sequences (questions and context) and output text sequences (answers) using a vocabulary of subword tokens.

T5 undergoes training using examples in a text-to-text format: (X, Y) where X represents the input text consisting of a question and context, and Y represents the desired output text, which is the answer.

The model is trained to minimize the negative log-likelihood of the target output given the input:

$$L(\theta) = -\frac{1}{N}\sum_{i=1}^{N} \log P(Y_i|X_i, \theta) \tag{1}$$

where θ represents the model parameters, N is the number of training examples, and $P(Y_i|X_i, \theta)$ is the probability assigned by the model to the target output given the input in (1).

The encoding process includes an input text X which contains a question and context, T5 tokenizes and encodes it into a sequence of input tokens $X = (x_1, x_2, \ldots, x_{|X|})$. The output sequence $Y = (y_1, y_2, \ldots, y_{|Y|})$ is generated autoregressively by the model one token at a time, conditioned on the input sequence X:

$$P(y_t|y_{<t}, X) = \text{softmax}(f_{\text{dec}}(y_{<t}, \text{enc}(X))) \tag{2}$$

where f_{dec} is the decoder function, enc is the encoder function, and $y_{<t}$ denotes the tokens generated before time step t in (2) and y_t in (3) is a greedy decoding strategy, selecting the highest probability token at each time step:

$$y_t = \arg\max_{y} P(y_t|y_{<t}, X) \tag{3}$$

T5 is evaluated using the negative log-likelihood loss function in (4) similar to training:

$$L_{\text{eval}} = -\frac{1}{N_{\text{eval}}}\sum_{i=1}^{N_{\text{eval}}} \log P(Y_i|X_i, \theta) \tag{4}$$

where N_{eval} is the number of examples in the evaluation dataset.

T5 signifies a significant and transformative change in NLP by considering all tasks as challenges of generating text. T5's adaptability, ability to transfer knowledge, and scalable structure make it a promising method for constructing precise and adaptable question-answering systems.

F. SIMILARITY ANALYSIS

To analyse the similarity between the topics obtained from LDA and T5 generated answers, Cosine Similarity method was employed. It is a metric used in text processing to determine the similarity between two non-zero vectors in an inner product space. It does this by calculating the cosine of the angle between the vectors. The scale of this measure spans from 0 to 1, with a value of 1 indicating a perfect match. Park et al. [37] were the first to use cosine similarity into traditional classifiers such as SVM, CNN and MNB in order to improve their performance. Significantly, the cosine of 0 is

Algorithm 1 Proposed Algorithm for Descriptive Answer Evaluation

Require: Exam questions Q, Student answers A_{student}, Ideal answers A_{ideal}, Weights: w_{LDA}, w_{T5}

Ensure: Evaluated scores for student answers

1: Perform data preprocessing on A_{student} and A_{ideal}

2: Apply LDA to A_{student} and A_{ideal} to extract thematic topics and represent each answer as a distribution over the identified topics
- Calculate cosine similarity ($\text{cos\_sim}_{\text{LDA}}$) between the topic distributions of A_{student} and A_{ideal} using:
$$\text{cos\_sim}_{\text{LDA}} = \frac{S_{\text{topic\_dist}} \cdot I_{\text{topic\_dist}}}{\|S_{\text{topic\_dist}}\| \cdot \|I_{\text{topic\_dist}}\|}$$

3: Fine-tune the T5 model on Q, A_{student}, and A_{ideal}

4: Use the fine-tuned T5 model to generate answers for A_{student} and A_{ideal}
- Compute SBERT embeddings (E_{student}, E_{ideal}) for the generated answers
- Calculate cosine similarity ($\text{cos\_sim}_{\text{SBERT}}$) between E_{student} and E_{ideal} using:
$$\text{cos\_sim}_{\text{SBERT}} = \frac{E_{\text{student}} \cdot E_{\text{ideal}}}{\|E_{\text{student}}\| \cdot \|E_{\text{ideal}}\|}$$

5: Aggregation of Scores:
- Compute the final score (final_score) as the weighted sum of $\text{cos\_sim}_{\text{LDA}}$ and $\text{cos\_sim}_{\text{SBERT}}$ using:

$\text{final\_score} = (w_{\text{LDA}} \cdot \text{cos\_sim}_{\text{LDA}}) + (w_{\text{T5}} \cdot \text{cos\_sim}_{\text{SBERT}})$

6: Repeat steps 2-6 for all student answers

7: **return** Evaluated scores for student answers

equal to 1, however for any other angle inside the interval, it remains less than 1.

Cosine Similarity measures similarity between the topics generated for student answers and ideal answers after applying LDA

$$\text{cos\_sim}_{\text{lda}} = \frac{(S_{\text{topic\_dist}} \cdot I_{\text{topic\_dist}})}{(\|S_{\text{topic\_dist}}\| \cdot \|I_{\text{topic\_dist}}\|)} \quad (5)$$

$$\text{cos\_sim}_{\text{sbert}} = \frac{(E_{\text{student}} \cdot E_{\text{ideal}})}{(\|E_{\text{student}}\| \cdot \|E_{\text{ideal}}\|)} \quad (6)$$

where in (5) and (6)

- $S_{\text{topic\_dist}}$: topic distribution of the student answer after LDA
- $I_{\text{topic\_dist}}$: topic distribution of the ideal answer after LDA
- E_{student}: Embedding of the student answer generated by T5
- E_{ideal}: Embedding of the ideal answer generated by T5

G. SENTENCE EMBEDDING

Sentence-BERT [38] is an advanced language model that use Siamese BERT networks for transforming sentences into embeddings. SBERT is designed to compare sentences or paragraphs semantically. It encodes phrase semantics into fixed-size vectors to bring semantically related sentences closer together in the embedding space. T5 measures semantic knowledge while generating text, while SBERT measures semantic similarity between sentences or sections. SBERT following T5 can refine the evaluation process by explicitly examining the student's generated answer's semantic content similarity to the ideal answer. Applying SBERT produces embeddings for T5 generated student answers, $E_{\text{student}}$and ideal answers, E_{ideal}

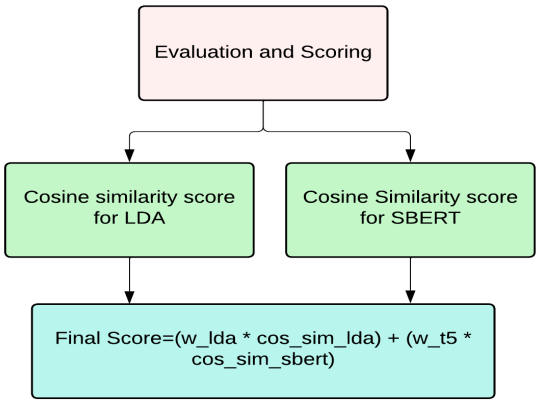

FIGURE 2. Evaluation and scoring module.

H. EVALUATION AND SCORING MODULE

This module computes the cosine similarity of the topics generated by employing LDA on both Student answers and Ideal answers. Also, similarity between encoded sentences of T5 generated student answers and ideal answers. The final score is generated by using the formula as:

$$\text{final\_score} = (w_{\text{lda}} \cdot \text{cos\_sim}_{\text{lda}}) + (w_{\text{t5}} \cdot \text{cos\_sim}_{\text{sbert}}) \quad (7)$$

In (7), w_{lda} is the weight assigned to the LDA model and w_{t5} is the weight assigned to the T5 model. Assigning weights to models when calculating the final score in a hybrid model is important for balancing contributions, optimizing performance, providing flexibility, mitigating biases, and enhancing robustness.

Figure 3 depicts the entire flow process of the proposed hybrid model that helps in the evaluation of student descriptive answers. The process involves the following steps:

- The process begins with the start of evaluation.
- Preprocessing is performed to prepare the data for LDA and T5 model.
- LDA Topic Modeling is applied to extract thematic topics from student and ideal answers.
- Cosine similarity is computed for the LDA topic distributions.
- T5 QA model is applied to extract answers from student answers and ideal answers.
- SBert model is applied on T5 generated answers to obtain Sentence embeddings.
- Cosine similarity is computed for the SBert embeddings of student and ideal answers.
- The final score is computed by aggregating the scores from LDA and T5.

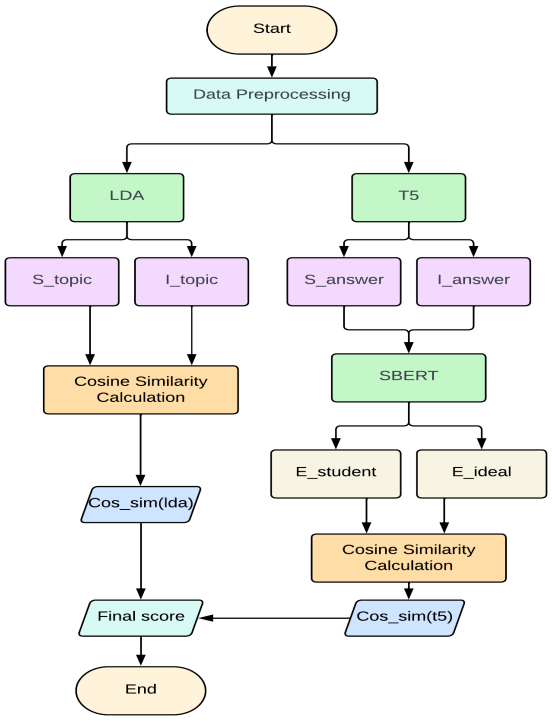

FIGURE 3. Flow process of the proposed hybrid model for evaluating student descriptive answers.

- The evaluation process ends after all student answers have been evaluated.

IV. EXPERIMENTAL RESULTS

A. EXPERIMENTAL SETUP

The configuration used for this experiment involves a Google Colab Python notebook executed with 12 GB RAM and an SSD of 512GB. GPU is not turned on for this. The data is partitioned in a ratio of 80:20 into training and validation sets. The testing set is obtained by considering a separate set of answers from 10 students for 5 questions. Table 3 shows the sizes of training, validation and testing sets. The results were obtained by using the proposed hybrid model using LDA and T5 incorporating Cosine similarity method.

TABLE 3. Data split configuration.

Training set	Validation set	Testing set
7200 instances	1800 instances	50 instances

B. EVALUATION METRICS

The metrics considered to evaluate this model are accuracy, precision, recall, f1-score and the scores generated by the proposed model. Though Accuracy may not be directly applicable since the model does not classify inputs into discrete categories as traditional classifiers do. However, the concept of a confusion matrix(cm) can be adapted to evaluate the performance of the model in a similar manner by categorizing True Positive(tp'),True Negative(tn'),False Positive(fp'),False Negative(fn') values. In this scenario, tp' refers to cases where both thematic coverage and semantic similarity are high, indicating accurate evaluation, fp' are the cases where thematic coverage is high but semantic similarity is low, indicating inaccurate evaluation, fn' denotes cases where thematic coverage is low but semantic similarity is high, indicating inaccurate evaluation. tn' are the cases where both thematic coverage and semantic similarity are low. Table 4 depicts the detailed metrics used for evaluation.

TABLE 4. Summary of evaluation metrics.

Metric	Description	Formula
Accuracy	Quantifies the degree of correctness in the predictions made by a model.	$\frac{tp'+tn'}{tp'+fp'+tn'+fn'}$
Precision	Computes the ratio of true positive predictions to the total number of positive predictions generated by the model.	$\frac{tp'}{tp'+fp'}$
Recall	Computes the ratio of correctly predicted positive cases to the total number of genuine positive cases in the dataset.	$\frac{tp'}{tp'+fn'}$
F1-score	The harmonic mean of recall and precision, offering an optimal blend of recall and precision for unified evaluation of the model's effectiveness.	$\frac{2*Precision*Recall}{Precision+Recall}$

1) COMPUTED SCORES BY THE MODEL

In addition to the traditional evaluation metrics, the hybrid model computes thematic coverage scores, semantic similarity scores, and overall evaluation scores for each student answer. The aforementioned scores hold significance in comprehending the model's efficacy in thoroughly evaluating student responses.

C. RESULTS

The process of assessing student's descriptive answers involves the use of the proposed model to derive the final outcome. LDA model produces topics for both the ideal answer, the student answer, and then calculates their similarity score. This assesses the extent to which student replies address the specific themes. Table 5 displays the similarity scores of the topics obtained for 10 student replies that were used for testing.

Once the scores from the LDA model have been obtained, the subsequent step is to evaluate the student's level of semantic comprehension. The process involves utilising the T5 model that provide answers for the questions. The generated answers are processed using an SentenceBERT embedding technique, which produces embeddings. A similarity score is then generated based on these embeddings. The similarity scores generated are displayed in Table 6, which pertains to 10 students.

The final scores are calculated by summing up the weighted scores generated by LDA and T5 models. The final

TABLE 5. Similarity scores prediction after applying LDA.

Student #	Question 1	Question 2	Question 3	Question 4	Question 5
s1	0.85	0.90	0.88	0.92	0.89
s2	0.80	0.82	0.85	0.79	0.83
s3	0.92	0.88	0.90	0.87	0.91
s4	0.78	0.81	0.77	0.83	0.80
s5	0.96	0.85	0.83	0.89	0.87
s6	0.9	0.91	0.95	0.9	0.93
s7	0.88	0.85	0.82	0.93	0.95
s8	0.91	0.97	0.96	0.99	0.99
s9	0.99	0.99	0.98	0.97	0.99
s10	1	1	0.92	0.98	0.97

TABLE 6. Similarity scores prediction after applying T5.

Student #	Question 1	Question 2	Question 3	Question 4	Question 5
s1	0.9	0.91	0.9	0.9	0.89
s2	0.82	0.79	0.84	0.81	0.80
s3	0.89	0.87	0.88	0.90	0.92
s4	0.75	0.76	0.74	0.77	0.78
s5	0.91	0.92	0.90	0.93	0.95
s6	1	1	1	0.9	0.89
s7	0.6	0.62	0.5	0.52	0.4
s8	0.99	0.99	0.98	0.99	0.9
s9	0.5	0.6	0.4	0.73	0.7
s10	0.8	0.87	0.84	0.9	0.89

score assesses both the thematic and semantic understanding of the student. The final score is obtained using the equation (7). Table 7 shows the final scores generated by the model. The weights for the models are necessary for the model's optimized performance in generating scores. Assigning weights to models involves a combination of empirical observation, domain expertise, experimentation, validation, and fine-tuning. It's an iterative process aimed at optimizing the performance of the hybrid model and ensuring that it effectively captures the desired aspects of the data for evaluation.

Assigning equal weights to the LDA and SBERT scores is a logical starting point, supported by the complementary strengths of thematic coverage and semantic understanding. This balanced approach ensures a comprehensive evaluation of student answers, taking into account both what topics are covered and how well they are articulated. Equal weighting helps in maintaining consistency and fairness across diverse responses. Students who provide comprehensive coverage of topics but might lack in-depth semantic details, or vice versa, are both evaluated fairly. This approach mitigates the risk of biased grading that could arise from emphasizing one aspect over the other. Empirical validation and further research can refine these weights, but the initial assumption of equal weighting is justified by the balanced contributions each method offers to the overall assessment. For the proposed model, as both the models play an important role in generating scores, weight of 0.5 was assigned equally to both the models.

Hence as illustrated in (7),

$\text{final\_score} = (w_{\text{lda}} \cdot \cos\_\text{sim}_{\text{lda}}) + (w_{\text{t5}} \cdot \cos\_\text{sim}_{\text{sbert}})$

For S1, final_score = (0.5 * 4.44) + (0.5 * 4.5) = 4.47. Likewise final score for all the students is generated.

TABLE 7. Final score prediction by the proposed model.

Student #	Final Score
s1	4.47
s2	4.075
s3	4.47
s4	3.91
s5	4.45
s6	4.69
s7	3.52
s8	4.83
s9	3.92
s10	4.58

As discussed earlier these answers were annotated and given scores by different human evaluators. Table 8 shows the comparison between model generated scores and human scores. Their differences can also be seen in the table 8. To facilitate comparability, the final results are multiplied by a value of 10, as each question is worth 10 marks and the annotators assigned scores on a scale of 10 for each question.

TABLE 8. Comparison of final scores with human scores.

Student #	Final Score	Human Score	Variation
s1	44.7	45	0.3
s2	40.7	40	0.7
s3	44.7	44	0.7
s4	39.1	38	1.1
s5	44.5	45	0.5
s6	46.9	45	1.9
s7	35.2	28	7.2
s8	48.3	49	0.7
s9	39.2	38	1.2
s10	45.8	44	1.8

As seen in the table the variation between final score and human score is only below 2.0 which is a very good score. Out of 10 students 9 student's scores are very near to the human evaluated scores. This tells that the answers given by student covers almost all the topics and content same as in the ideal answer for that question. The Accuracy, Precision, Recall, F1-score values for the proposed model are given in table 9. The model achieved a training accuracy of 95%, validation accuracy of 92% and the testing accuracy of 91% Different QA models were used and the accuracies

TABLE 9. Performance metrics of the proposed model.

	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Training	95	94	93	93.3
Validation	92	91	89	90
Testing	91	91	92	91

were compared based on their generated final scores for the student answers. As illustrated in Table 10, LDA is presented as the topic modeling model in conjunction with other QA models, including Bidirectional Encoder Representations from Transformers (BERT), Robustly optimized BERT approach (RoBERTa), and Distilled BERT (DistilBERT). BERT, introduced by Google researchers in 2018, utilizes

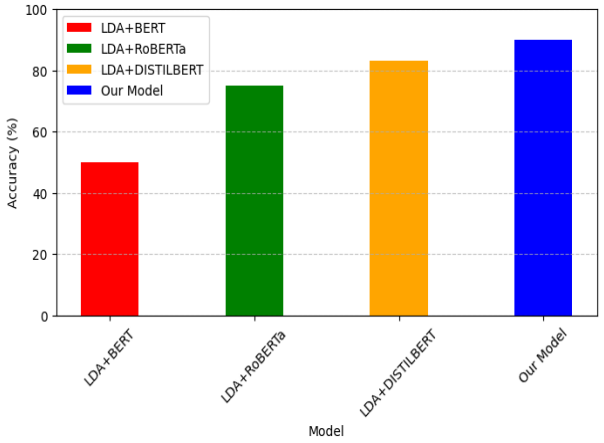

FIGURE 4. Comparison of accuracy for different models.

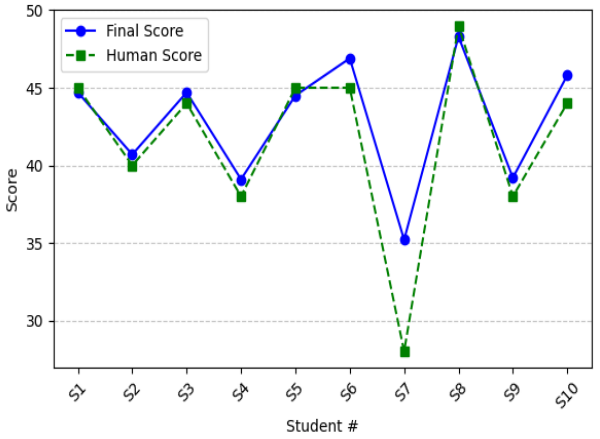

FIGURE 5. Comparison of predicted final scores and human assessed scores.

the Transformer architecture to pre-train deep bidirectional representations from large unlabeled text datasets. It can then be fine-tuned for various NLP tasks. RoBERTa, an enhanced version of BERT developed by Facebook AI, surpasses BERT by training on more extensive datasets for longer durations and with larger batch sizes, among other enhancements. Distilled BERT, a creation of researchers at Hugging Face, is a compact and faster alternative to BERT that maintains most of its performance while employing fewer parameters. This makes Distilled BERT more lightweight and suitable for deployment in environments with limited computational resources or applications with constrained resources.

Table 10 displays the accuracy comparison of different models in combination with LDA along with the proposed hybrid model that is used to evaluate student descriptive answers. As it is seen, BERT achieved only 50% accuracy, RoBERTa performed better than BERT and achieved an accuracy of 75%, DistilBERT achieved 83% and the proposed model outperformed the other models by achieving an accuracy of 91%.

TABLE 10. Comparison of model accuracy.

Model	Accuracy (%)
LDA+BERT	50
LDA+RoBERTa	75
LDA+DistilBERT	83
Our Model	91

Figure 4 illustrates the compared accuracies for different models. As discussed above, it is evident from figure 4 that the proposed hybrid model is successful in evaluating student descriptive answers more accurately when compared to other models.

The variation between model generated scores and human scores of 10 students can be seen in figure 5. As it shows, only student#7 s7 score is far from the human score. s7 scored 35.2 as generated by the model, while human evaluator gave a score of 28 only making a difference of 7.2 marks.

D. DISCUSSION

To automatically evaluate student's descriptive answers, various models have been discussed in prior studies. However, these studies typically rely on keyword evaluation, grammar consideration, syntactical analysis, etc. The proposed hybrid model integrates LDA for thematic coverage and T5 for semantic understanding, combining statistical and machine learning approaches for evaluation. In [39], the authors proposed a fusion of fuzzy ontology with Latent Semantic Analysis (LSA), involving the retrieval of syntax and semantic features. They achieved an accuracy of 0.77 using a multiple linear regression model, indicating a combination of statistical and rule-based methods.

In another study [40], the authors implemented an approach based on text mining for short answer grading, comparing model answers with student responses by calculating sentence distances. They achieved a correlation of 0.81 between student and model answers, indicating a primarily rule-based approach based on completeness and vocabulary matching.

While these approaches focus on specific features or rule-based comparisons, the proposed hybrid model combines multiple methods to capture thematic relevance, semantic understanding, and overall answer quality. It offers a more versatile and comprehensive approach to evaluating student answers, leveraging the strengths of both statistical and machine learning techniques for more accurate assessment.

The proposed hybrid model has significant practical implications for automating the evaluation of student answers in educational settings. By enhancing efficiency, consistency, scalability, and providing personalized feedback, it has the potential to revolutionize the assessment process, benefiting both educators and students. However, successful deployment requires careful consideration of implementation challenges and ongoing efforts to refine and improve the model's performance.

The proposed DAES model offers several notable advantages. By integrating topic modelling and QA models,

it provides a comprehensive evaluation of student answers, capturing both thematic coverage and semantic understanding. LDA helps identify key topics and themes, while T5 ensures the semantic accuracy and relevance of the answers. This dual approach has demonstrated high accuracy of 91% in evaluating student answers, significantly improving the reliability of automated grading systems. Furthermore, the automated nature of the model allows for the efficient evaluation of large volumes of student answers, making it scalable for use in large classrooms or online courses, thereby reducing the time and effort required for manual grading. The model also minimizes subjective biases and inconsistencies inherent in human grading by standardizing the evaluation process, ensuring fair and consistent assessment of all student responses. Additionally, by analyzing both the content and context of student answers, the model can provide detailed and constructive feedback, helping students understand their mistakes and improve their learning outcomes.

Despite its advantages, the DAES model also has several limitations. The integration of LDA and T5 introduces complexity in terms of implementation and maintenance, requiring a thorough understanding of both techniques. The model's performance is highly dependent on the quality and quantity of training data; inadequate or biased training data can lead to poor evaluation results, making it crucial to ensure a diverse and representative dataset. Training and fine-tuning deep learning models like T5 require significant computational resources, which might be a limitation for institutions with limited access to high-performance computing infrastructure. Although the model can be fine-tuned for specific domains, this process requires additional effort and expertise, and the model might not perform equally well across all subjects without domain-specific adjustments.

V. CONCLUSION

A novel hybrid model, DAES is proposed in this research for the automatic evaluation of student descriptive answers, representing a significant advancement in educational assessment methodology. By integrating LDA for thematic coverage and T5 for semantic understanding, a versatile and comprehensive approach has been developed which is capable of accurately evaluating student responses across diverse subjects and topics. Through rigorous experimentation and validation, an accuracy of 91%, precision of 91%, recall of 92%, and an F1-score of 91% was achieved, demonstrating the effectiveness and reliability of our model in assessing student performance.

The practical implications of our proposed approach have a wide-ranging impact. The capacity to expedite the evaluation process, guarantee consistency and objectivity, and deliver personalized feedback provides educators with a potent tool for improving teaching and learning results. Furthermore, its scalability and adaptability make it suitable for deployment in a wide range of educational settings, from classrooms to online learning platforms.

While our model has demonstrated strong performance, the importance of ongoing refinement and improvement can be recognized. Future research efforts will focus on expanding the training dataset and exploring additional modalities, which may include consideration of diagrams, mathematical equations, programming code, along with text in answer scripts, to further enhance its accuracy and robustness. A wide range of educational domains and subjects, including but not limited to mathematics, social studies, science, language arts can be implemented for evaluation.

REFERENCES

[1] J. Wang and Y. Dong, "Measurement of text similarity: A survey," *Information*, vol. 11, no. 9, p. 421, Aug. 2020.
[2] M. Han, X. Zhang, X. Yuan, J. Jiang, W. Yun, and C. Gao, "A survey on the techniques, applications, and performance of short text semantic similarity," *Concurrency Comput., Pract. Exper.*, vol. 33, no. 5, p. e5971, Mar. 2021.
[3] M. S. M. Patil and M. S. Patil, "Evaluating student descriptive answers using natural language processing," *Int. J. Eng. Res. Technol.*, vol. 3, no. 3, pp. 1716–1718, 2014.
[4] P. Patil, S. Patil, V. Miniyar, and A. Bandal, "Subjective answer evaluation using machine learning," *Int. J. Pure Appl. Math.*, vol. 118, no. 24, p. 113, 2018.
[5] J. Muangprathub, S. Kajornkasirat, and A. Wanichsombat, "Document plagiarism detection using a new concept similarity in formal concept analysis," *J. Appl. Math.*, vol. 2021, pp. 1–10, Mar. 2021.
[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] M. Kusner, Y. Sun, N. Kolkin, and K. Weinberger, "From word embeddings to document distances," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 957–966.
[8] C. Xia, T. He, W. Li, Z. Qin, and Z. Zou, "Similarity analysis of law documents based on word2vec," in *Proc. IEEE 19th Int. Conf. Softw. Qual., Rel. Secur. Companion (QRS-C)*, Jul. 2019, pp. 354–357.
[9] H. Mittal and M. S. Devi, "Subjective evaluation: A comparison of several statistical techniques," *Appl. Artif. Intell.*, vol. 32, no. 1, pp. 85–95, Jan. 2018.
[10] 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.
[11] G. Srivastava, P. K. R. Maddikunta, and T. R. Gadekallu, "A two-stage text feature selection algorithm for improving text classification," China Med. Univ., Taiwan, Tech. Rep. 137, 2021.
[12] H. Mangassarian and H. Artail, "A general framework for subjective information extraction from unstructured English text," *Data Knowl. Eng.*, vol. 62, no. 2, pp. 352–367, Aug. 2007.
[13] B. Oral, E. Emekligil, S. Arslan, and G. Eryiğit, "Information extraction from text intensive and visually rich banking documents," *Inf. Process. Manag.*, vol. 57, no. 6, Nov. 2020, Art. no. 102361.
[14] V. Nandini and P. Uma Maheswari, "Automatic assessment of descriptive answers in online examination system using semantic relational features," *J. Supercomput.*, vol. 76, no. 6, pp. 4430–4448, Jun. 2020.
[15] H. Tuan Nguyen, C. Tuan Nguyen, H. Oka, T. Ishioka, and M. Nakagawa, "Handwriting recognition and automatic scoring for descriptive answers in Japanese language tests," 2022, *arXiv:2201.03215*.
[16] Y. Wu, A. Henriksson, J. Nouri, M. Duneld, and X. Li, "Beyond benchmarks: Spotting key topical sentences while improving automated essay scoring performance with topic-aware BERT," *Electronics*, vol. 12, no. 1, p. 150, Dec. 2022.
[17] M. F. Bashir, H. Arshad, A. R. Javed, N. Kryvinska, and S. S. Band, "Subjective answers evaluation using machine learning and natural language processing," *IEEE Access*, vol. 9, pp. 158972–158983, 2021.
[18] R. S. Wagh and D. Anand, "Legal document similarity: A multi-criteria decision-making perspective," *PeerJ Comput. Sci.*, vol. 6, p. e262, Mar. 2020.
[19] M. Alian and A. Awajan, "Factors affecting sentence similarity and paraphrasing identification," *Int. J. Speech Technol.*, vol. 23, no. 4, pp. 851–859, Dec. 2020.

[20] G. Jain and D. K. Lobiyal, "Conceptual graphs based approach for subjective answers evaluation," *Int. J. Conceptual Struct. Smart Appl.*, vol. 5, no. 2, pp. 1–21, Jul. 2017.
[21] M. Montes-Y-Gómez, A. López-López, and A. Gelbukh, "Information retrieval with conceptual graph matching," in *Proc. Int. Conf. Database Expert Syst. Appl.*, vol. 1873, Jan. 2000, pp. 312–321.
[22] V. Bahel and A. Thomas, "Text similarity analysis for evaluation of descriptive answers," 2021, *arXiv:2105.02935*.
[23] H. Zhang and D. Litman, "Automated topical component extraction using neural network attention scores from source-based essay scoring," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, Art. no. 85698584.
[24] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Jan. 2003.
[25] C. K. Carter and R. Kohn, "On Gibbs sampling for state space models," *Biometrika*, vol. 81, no. 3, pp. 541–553, Aug. 1994.
[26] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, pp. 1–67, Jan. 2020.
[27] G. Grefenstette, "Tokenization," in *Syntactic Wordclass Tagging*. Berlin, Germany: Springer, 1999, pp. 117–133.
[28] K. Sirts and K. Peekman, "Evaluating sentence segmentation and word Tokenization systems on Estonian web texts," in *Proc. 9th Int. Conf. Baltic (HLT)*, vol. 328, U. Andrius, V. Jurgita, K. Jolantai, and K. Danguole, Eds. Kaunas, Lithuania: IOS Press, Sep. 2020, pp. 174–181.
[29] A. Schofield, M. Magnusson, and D. M. Mimno, "Pulling out the stops: Rethinking stopword removal for topic models," in *Proc. 15th Conf. Eur. Chapter Assoc. Comput. Linguistics (EACL)*, vol. 2, M. Lapata, P. Blunsom, and A. Koller, Eds. Valencia, Spain: Association for Computational Linguistics, 2017, pp. 432–436.
[30] M. Çagatayli and E. Çelebi, "The effect of stemming and stop-word removal on automatic text classification in Turkish language," in *Proc. 22nd Int. Conf. Neural Inf. Process. (ICONIP)* (Lecture Notes Computer Science), vol. 9489, S. Arik, T. Huang, W. K. Lai, and Q. Liu, Eds. Istanbul, Turkey: Springer, 2015, pp. 168–176.
[31] M. Divyapushpalakshmi and R. Ramalakshmi, "An efficient sentimental analysis using hybrid deep learning and optimization technique for Twitter using parts of speech (POS) tagging," *Int. J. Speech Technol.*, vol. 24, no. 2, pp. 329–339, Jun. 2021.
[32] F. Camastra and G. Razi, "Italian text categorization with lemmatization and support vector machines," in *Neural Approaches to Dynamics of Signal Exchanges* (Smart Innovation, Systems and Technologies), vol. 151, A. Esposito, M. Faúndez-Zanuy, F. C. Morabito, and E. Pasero, Eds. New York, NY, USA: Springer, 2020, pp. 47–54.
[33] A. Jabbar, S. Iqbal, M. I. Tamimy, S. Hussain, and A. Akhunzada, "Empirical evaluation and study of text stemming algorithms," *Artif. Intell. Rev.*, vol. 53, no. 8, pp. 5559–5588, Dec. 2020.
[34] C. Sammut and G. I. Webb, "TF-IDF," in *Encyclopedia of Machine Learning*, C. Sammut and G. I. Webb, Eds. Berlin, Germany: Springer, 2010, pp. 986–987.
[35] L. Havrlant and V. Kreinovich, "A simple probabilistic explanation of term frequency-inverse document frequency (TF-IDF) heuristic (and variations motivated by this explanation)," *Int. J. Gen. Syst.*, vol. 46, no. 1, pp. 27–36, Mar. 2017.
[36] A. Thakkar and K. Chaudhari, "Predicting stock trend using an integrated term frequency–inverse document frequency-based feature weight matrix with neural networks," *Appl. Soft Comput.*, vol. 96, Nov. 2020, Art. no. 106684.
[37] K. Park, J. S. Hong, and W. Kim, "A methodology combining cosine similarity with classifier for text classification," *Appl. Artif. Intell.*, vol. 34, no. 5, pp. 396–411, Apr. 2020.
[38] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," 2019, *arXiv:1908.10084*.
[39] S. M. Darwish and S. K. Mohamed, "Automated essay evaluation based on fusion of fuzzy ontology and latent semantic analysis," in *Proc. Int. Conf. Adv. Mach. Learn. Technol. Appl.*, 2020, pp. 566–575.
[40] N. Süzen, A. N. Gorban, J. Levesley, and E. M. Mirkes, "Automatic short answer grading and feedback using text mining methods," *Proc. Comput. Sci.*, vol. 169, pp. 726–743, Jan. 2020.
[41] P. Rajpurkar, J. Zhang, K. Lopyrev, and P. Liang, "SQuAD: 100,000+ questions for machine comprehension of text," 2016, *arXiv:1606.05250*.

LALITHA MANASA CHANDRAPATI received the B.Tech. degree in computer science and information technology discipline from Jawaharlal Nehru Technological University and the M.Tech. degree in computer science and engineering from Vignan's Foundation for Science, Technology and Research University, Vadlamudi. She is currently pursuing the Ph.D. degree and a full time Research Scholar with the School of Computer Science and Engineering, VIT-AP University, Amaravati. Her research interests include natural language processing, deep learning, and machine learning.

CH. KOTESWARA RAO (Member, IEEE) received the B.Tech. degree from Acharya Nagarjuna University, in 2004, the M.Tech. degree in CSE from JNTU Kakinada Campus, in 2009, and the Ph.D. degree in computer science and engineering from NIT Tiruchirappalli, Tamil Nadu, India, in 2021. He is currently an Assistant Professor with the School of Computer Science and Engineering, VIT-AP University. He has more than 17 years of experience, including 11 years of teaching and six years of research. He has published several research papers in reputed journals and conferences. His research interests include cryptography, information security, secure algorithms, and cyber security. He is a member of CRSI life time.