e-ISSN:2582-7219

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

Volume 7, Issue 13, April 2024

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.521

International Journal of Multidisciplinary Research in Science, Engineering and Technology (IJMRSET)

| ISSN: 2582-7219 | www.ijmrset.com | Impact Factor: 7.521| | Monthly Peer Reviewed & Refereed Journal |

|| Volume 7, Issue 13, April 2024 ||

International Conference on Intelligent Computing & Information Technology (ICIT-24)

Organized by

Erode Sengunthar Engineering College, Erode, Tamilnadu, India

Automated Answer Evaluation Based on Deep Learning

Mrs.S.Mahalakshmi[1] Ms.R.Dhanya[2], Ms.B.Maheswari [3], Ms.A.Nancy [4]

Assistant Professor, Dept. of C.S.E., Shree Venkateshwara Hi-Tech Engineering College, Gobichettipalayam, Erode

Tamil Nadu India[1]

Student, Dept. of C.S.E., Shree Venkateshwara Hi-Tech Engineering College, Gobichettipalayam, Erode

Tamil Nadu India[2,3,4]

ABSTRACT: Automated answer evaluation based on deep learning is a novel approach to streamline the grading process in educational settings. This project aims to develop a system capable of assessing students' answers to open-ended questions accurately and efficiently. Leveraging deep learning techniques, such as natural language processing and neural networks, the system will analyze text responses, identify key concepts, and assign scores based on predefined criteria. By automating the evaluation process, educators can save time, provide more timely feedback to students, and ensure consistency in grading. This paper discusses the architecture, implementation, and evaluation of the automated answer evaluation system, highlighting its potential to revolutionize the assessment process in education.

KEYWORDS: Automated assessment, Deep learning, Natural language processing, Neural networks, Answer grading, Educational technology, Machine learning, Evaluation system, Open- ended questions, Educational assessment.

I. INTRODUCTION

Optical character recognition (OCR) is the process of identifying individual letters and words within a digital image. It involves employing a classification algorithm to analyze each character and assemble them into coherent words. This method relies on algorithms that group similar words together, comparing the outcome with the expected text in the image. OCR technology converts images of printed, typed, or handwritten text into machine-encoded text, facilitating its extraction and manipulation by0020computers. For businesses, OCR is invaluable for quickly extracting and transforming text from scanned documents and images into a readable, editable, and searchable format. It enables computers to interpret written language much like the brain and eyes work together to comprehend text from images. Despite initial challenges, automated OCR systems like Tesseract have become widely available for integration into various programs, enhancing their functionality. Meanwhile, the traditional method of manually grading subjective responses, common in educational assessments, is increasingly impractical, particularly in the current remote working environment exacerbated by the pandemic. Although automated systems effectively evaluate multiple- choice or objective questions, they fall short in assessing subjective responses, which are crucial for understanding a student's comprehension and depth of understanding. Manual evaluation of such responses is time-consuming and labor-intensive, leading to inconsistencies in scoring and potentially affecting student performance. To address this issue, machine learning techniques, particularly deep learning algorithms, are being explored to automate the grading process, particularly in areas like natural language processing and image recognition. These algorithms can discern intricate patterns and correlations within data, making them well-suited for assessing subjective responses and providing accurate grades efficiently.

II. AUTOMATED ANSWER EVALUATION

Automated answer evaluation using deep learning represents a significant advancement in educational technology, offering the potential to streamline the grading process, provide instant feedback to students, and facilitate personalized learning experiences. By harnessing the power of neural networks with multiple layers, deep learning techniques enable computers to analyze and assess student responses with a level of sophistication that approaches human- like understanding. In this comprehensive approach, various neural network architectures, including recurrent neural networks (RNNs), convolutional neural networks (CNNs), and transformer models, play crucial roles in understanding and evaluating the content, context, and quality of student answers. RNNs, renowned for their ability to process sequences of data, are particularly effective for analyzing textual inputs like student responses. These networks can capture the sequential dependencies and relationships within sentences, enabling them to understand the flow of ideas and concepts presented by the students. By training on annotated datasets of student responses, RNNs learn to identify patterns of correctness, relevance, and coherence, providing valuable insights into the quality of the answers. Similarly, CNNs are instrumental in automated answer evaluation by extracting meaningful features from textual inputs. These networks excel at capturing local patterns in data, making them well-suited for identifying important phrases, keywords, or structural elements within student responses. By applying convolutional operations, CNNs can detect relevant information embedded in the text, facilitating accurate assessments of the content's appropriateness and relevance to the given question or prompt. Transformer models, such as BERT (Bidirectional Encoder Representations from Transformers), have emerged as powerful tools for natural language understanding tasks, including automated answer evaluation. These models leverage attention mechanisms to capture contextual information from both preceding and subsequent words in a sentence, enabling them to grasp nuanced meanings and infer the semantic coherence of student responses. By considering the global context of the text, transformer-based approaches can provide more nuanced evaluations, taking into account the broader context of the entire answer rather than focusing solely on individual words or phrases.

III. AUTOMATED SCRIPT ANALYZERS

Automated script analyzers represent a significant advancement in various fields, including software development, cybersecurity, and natural language processing. These tools utilize algorithms and techniques from machine learning and artificial intelligence to automatically analyze and assess the quality, correctness, and security aspects of scripts, code snippets, or textual documents.In software development, automated script analyzers play a crucial role in ensuring code quality and adherence to coding standards. These tools can detect common programming errors, identify potential performance bottlenecks, and suggest improvements to enhance the efficiency and maintainability of the codebase. By automatically reviewing code submissions, developers can save time on manual code reviews and focus their efforts on more critical tasks.In cybersecurity, automated script analyzers help detect and mitigate security vulnerabilities in scripts and codebases. These tools can identify potential security loopholes, such as SQL injection, cross-site scripting (XSS), or buffer overflow vulnerabilities, and provide recommendations for remediation. By proactively identifying and addressing security issues, organizations can strengthen their defenses against cyber threats and protect sensitive data from unauthorized access or exploitation.Moreover, automated script analyzers find applications in natural language processing tasks, such as sentiment analysis, text classification, or entity recognition. These tools can analyze and extract insights from textual documents, social media posts, or customer reviews, enabling organizations to gain valuable insights into customer preferences, market trends,

and public opinion. By automating the analysis of large volumes of text data, businesses can make informed decisions and drive strategic initiatives more effectively.The development of automated script analyzers involves the use of various machine learning algorithms and techniques, including supervised learning, unsupervised learning, and deep learning. Supervised learning algorithms are trained on labeled datasets containing examples of scripts or text documents along with their corresponding labels or classifications. These algorithms learn to identify patterns and relationships between input features and output labels, enabling them to make predictions on unseen data.Unsupervised learning algorithms, on the other hand, do not require labeled data for training and instead focus on discovering hidden patterns or structures within the input data. These algorithms are often used for tasks such as clustering similar scripts or identifying anomalies in codebases.Deep learning techniques, such as recurrent neural networks (RNNs), convolutional neural networks (CNNs), and transformer models, have shown remarkable performance in automated script analysis tasks. These models can capture complex relationships and dependencies within textual data, enabling them to generate accurate predictions and insights.

IV. DEEP LEARNING

Deep learning is a subset of machine learning that employs artificial neural networks with multiple layers to model and extract patterns from complex data. These networks, inspired by the structure and function of the human brain, are capable of learning hierarchical representations of data, enabling them to solve intricate tasks such as image recognition, natural language processing, and speech recognition.Key components of deep learning include neural networks, which consist of interconnected nodes organized into layers. These models typically comprise input, hidden, and output layers, with each layer performing specific computations. Activation functions introduce non- linearities into the network, enabling it to learn complex relationships between inputs and outputs. Training algorithms such as stochastic gradient descent optimize the network's parameters by minimizing a loss function that quantifies the disparity between predicted and actual outputs.Deep learning finds applications across diverse domains. In computer vision, convolutional neural networks have revolutionized tasks like image classification and object detection. In natural language processing, recurrent neural networks and transformer models have significantly advanced tasks such as language translation and sentiment analysis. Speech recognition systems leverage deep learning algorithms for accurate transcription of spoken language.In healthcare, deep learning models are employed in medical image analysis, disease diagnosis, drug discovery, and personalized treatment planning. Autonomous vehicles utilize deep learning for tasks like object detection, localization, and path planning.Despite its widespread applications, deep learning poses challenges. It requires large amounts of labeled data for training and can be computationally intensive. Interpreting deep learning models can be challenging, and they are susceptible to overfitting, where they memorize noise in the training data rather than learning generalizable patterns.The future of deep learning involves continued advancements in architectures, training algorithms, and interdisciplinary applications. Addressing ethical concerns related to privacy, bias, fairness, and accountability is crucial. There is also growing interest in developing techniques to enhance the interpretability and transparency of deep learning models.

V. RELATED WORK

The proposed model seeks to develop an Automatic Arabic Short Answer Grading (AASAG) system by employing semantic similarity techniques. The model utilizes Latent Semantic Analysis (LSA) to gauge the semantic likeness between the Student Answer (SA) and Model Answer (MA). It incorporates two weighting methods, namely local weighting schema and hybrid local and global weighting schema, to fill the cell values .Several approaches have been suggested for automated grading of short answers. In one study the authors introduced AR-ASAG, an Arabic Dataset for Automatic Short Answer Grading Evaluation. They also presented an Automatic Short Answer Grading method based on the COALS (Correlated Occurrence Analogue to Lexical Semantic) algorithm, which demonstrated promising outcomes for Arabic language assessments. This method was tested on the AR-ASAG dataset, consisting of 2133 pairs of (Model Answer, Student Answer) in various formats. Another study introduced an Automatic Arabic Essay Scoring (AAES) system utilizing the Vector Space Model (VSM) and Latent Semantic Indexing (LSI). This approach, applied to a single question with four model answers and 30 student responses, involved information retrieval techniques followed by VSM and LSI to measure similarity between student and instructor essays. In a different study an automatic Arabic essay grading (AAEG) system was proposed using Support Vector Machine (SVM), which extracted features from student and model answers and identified related words using AWN (Arabic WordNet). This model, tested on multiple languages and a Kaggle dataset with 40 questions and 120 model answers, demonstrated improved accuracy with AWAN integration. Additionally, an automatic scoring system for short Arabic texts was proposed utilizing a sentence embedding approach and tested on various datasets including AraScore and AR- ASAG. Authors in another study proposed an automatic scoring system for Arabic short answers using Longest Common Subsequence (LCS) and Arabic WordNet (AWAN). This model achieved a Root Mean Square Error (RMSE) value of 0.81 and a Pearson correlation r value of 0.94 on a dataset with 330 student answers. Lastly, a study presented an automatic grading system for Arabic short answer questions using an optimized deep learning model, specifically a hybrid LSTM (Long Short-Term Memory) and GWO (Grey Wolf Optimizer) model. Tested on a dataset gathered from science subjects in various schools, this LSTM-GWO model outperformed other methods such as SVM, Ngram, Word2vec, and MaLSTM in accuracy.

VI. EXISTING SYSTEM

Automated answer evaluation systems leveraging deep learning undergo a multi-stage process. Initially, textual answers are preprocessed, involving tasks such as cleaning and tokenization to ensure data uniformity and readability. Subsequently, feature extraction is employed to convert the text into numerical representations, often utilizing methodologies like word embeddings to capture semantic meanings effectively. Following this, the model training phase commences, wherein deep learning architectures such as recurrent neural networks (RNNs), convolutional neural networks (CNNs), or transformer models are deployed to discern intricate patterns within the textual data. These architectures are trained on large datasets to learn how to accurately assess the quality of answers. Finally, the system undergoes evaluation, where predicted scores are compared against ground truth labels to gauge the system's efficacy. This evaluation phase is crucial for assessing the system's performance and identifying areas for improvement. Overall, through these stages of preprocessing, feature extraction, model training, and evaluation, automated answer evaluation systems employing deep learning algorithms streamline the assessment process, offering scalable and efficient solutions for educational institutions and online learning platforms.

VII. Proposed System

A proposed automated answer evaluation system utilizing deep learning would integrate advanced natural language processing (NLP) techniques to enhance evaluation accuracy and efficiency. The system would begin by preprocessing textual answers, including tasks such as text normalization, spelling correction, and grammar checking, to ensure data cleanliness and coherence. Next, the system would leverage state-of-the-art deep learning models, such as transformer-based architectures like BERT or GPT, to extract high-level semantic features from the answers. These models have demonstrated exceptional performance in understanding context and semantics, making them ideal for assessing the quality of textual responses.Furthermore, the system would incorporate domain-specific knowledge to tailor the evaluation process according to the subject matter. For instance, in evaluating scientific or technical responses, the system could integrate domain-specific embeddings or ontologies to enhance its understanding of specialized terminology and concepts.To improve the system's adaptability and robustness, it would employ techniques such as ensemble learning, where multiple models are combined to make predictions, reducing the risk of overfitting and enhancing overall performance.Finally, continuous monitoring and feedback mechanisms would be integrated into the system to iteratively improve its evaluation capabilities over time. User feedback and system performance metrics would be used to fine-tune model parameters and update the system's knowledge base, ensuring ongoing optimization and effectiveness in assessing answers accurately.

VIII. About The RNN

Recurrent Neural Networks (RNNs) are a class of neural networks particularly well-suited for sequential data processing, making them a popular choice for tasks like automated answer evaluation. Unlike feedforward neural networks, which process input data independently, RNNs have connections that form a directed cycle, allowing them to maintain internal state and capture dependencies across time steps.In the context of answer evaluation, RNNs can analyze textual answers word by word, considering the sequential nature of language. This enables them to capture the contextual information and dependencies present in the answers. For example, when evaluating a paragraph-long response to a question, an RNN can analyze each word in the response while retaining information from previous words to understand the overall meaning and coherence.One common variant of RNNs is the Long Short-Term Memory (LSTM) network, which addresses the vanishing gradient problem encountered during training. LSTMs have mechanisms called gates that regulate the flow of information, allowing them to capture long-range dependencies more effectively than traditional RNNs.In automated answer evaluation systems, RNNs can be trained on labeled datasets containing pairs of student answers and corresponding scores. By learning from these examples, the RNNs can develop the ability to predict scores for new answers based on their content and structure. Additionally, RNNs can be integrated with other components such as attention mechanisms to focus on relevant parts of the answer during evaluation, further enhancing their performance. Overall, RNNs offer a powerful framework for automated answer evaluation by leveraging the sequential nature of textual data.

IX. Pre-Processing

Preprocessing for automated answer evaluation using deep learning encompasses several crucial steps to prepare textual data for effective modeling. Initially, tokenization breaks text into individual units, usually words or subwords. Lowercasing ensures uniformity by converting all characters to lowercase, avoiding distinctions based on case. Removing punctuation eliminates non-semantic symbols like periods or commas. Stopwords removal discards common, low- information words such as "and" or "the". Normalization further standardizes text by reducing words to their base or root forms, aiding in capturing semantic similarities. Padding and truncation ensure uniform sequence lengths, vital for neural network processing. Vectorization converts text tokens into numerical embeddings, representing semantic relationships and enhancing model comprehension. Additionally, data augmentation techniques like synonym replacement or paraphrasing can enhance training diversity and model robustness. Together, these preprocessing steps create a clean, standardized textual dataset ready for training deep learning models to accurately evaluate answers automatically, enabling efficient assessment and feedback mechanisms in various educational or evaluative contexts.

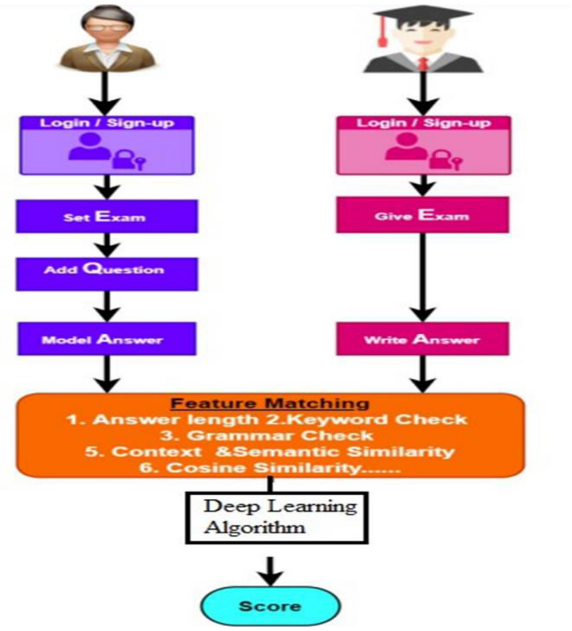
X. Data Preprocessing Module

The Data Preprocessing Module is responsible for preparing the input data for the automated answer evaluation system. It includes the following tasks: Text Cleaning and Normalization: Remove special characters, punctuation, and non-alphanumeric characters from the text. Convert text to lowercase to ensure consistency. Remove extra whitespaces and trim the text.Tokenization: Split the text into individual words or subwords. Create tokens to represent each word or subword separately.Padding or Truncation: Ensure all input sequences have the same length. Pad shorter sequences with zeros or truncate longer sequences.Word Embedding: Convert words into dense vectors using techniques like Word2Vec or GloVe Use pre-trained embeddings to capture semantic relationships between words. Handling Out-of-Vocabulary Words: Replace out-of-vocabulary words with a special token or unknown token.

XI. Evaluation And Scoring Module

Comparison: Compare student responses with model answers using similarity metrics such as cosine similarity or Levenshtein distance.Scoring: Assign scores or labels to student responses based on their similarity to model answers. Evaluation Metrics: Calculate evaluation metrics such as accuracy, precision, recall, and F1-score to measure the performance of the automated evaluation system.

XII. Architecture Diagram



XIII. Result

It is customary to administer a series of assessments to all students within an educational institution to assess their performance. Upon investigation, it became evident that while there are numerous techniques available for grading objective answers, there are considerably fewer methods for evaluating descriptive responses. When a professor assesses a descriptive answer, they typically focus on specific terms that aid in determining the accuracy of the response. Our system examines the document, employs optical character recognition to extract pertinent keywords, and subsequently utilizes cosine similarity to contrast these keywords with those provided by the user. Subsequently, our system will display a comparison of your responses based on this analysis.

XIV. Conclusion And Feature Work

An automated evaluation system plays a crucial role in assessing open-ended inquiries like short responses and essays, providing several advantages such as simplifying manual grading processes, saving time, effort, and resources, and ensuring fairness in evaluating students' responses. This paper introduces the concept of Automatic Arabic Short Answer Grading, utilizing Latent Semantic Analysis (LSA), a commonly used corpus-based similarity method. The system is applied to AR-ASAG, a limited Arabic dataset available to the public. Two experiments are conducted, employing different strategies for weighting data representation: local weighting and a hybrid approach that combines local and global weighting. The hybrid method, which integrates both local and global weights with LSA, surpasses the performance of the local weighting-based LSA, achieving an F1-score of 82.82% and an RMSE value of 0.798. Furthermore, the proposed weighting techniques demonstrate superior effectiveness compared to existing methods. Future research endeavors to improve the accuracy of the grading system, expand testing to encompass additional languages, and incorporate Arabic WordNet to refine the efficiency of scoring short answer questions.

References

[1] S. W. N. Cheung, S. C. Ng, and A. K. F. Lui, ‘‘A framework for effectively utilising human grading input in automated short answer grading,’’ Int.J. Mobile Learn. Organisation, vol. 16, no. 3, p. 266, 2022.

[2] W. H. Gomaa and A. A. Fahmy, ‘‘A survey of text similarity approaches,’’Int. J. Comput. Appl., vol. 13, no. 1, pp. 11–23, 2013.

[3] R. Mihalcea, C. Corley, and C. Strapparava, ‘‘Corpus-based and knowledge-based measures of text semantic similarity,’’ in Proc. AAAI, vol. 6, 2006, pp. 775–780.

[4] R. A. Farouk, M. H. Khafagy, M. Ali, K. Munir, and R. M. Badry,‘‘Arabic semantic similarity
approach for Farmers’ complaints,’’ Int. J. Adv. Comput. Sci. Appl., vol. 12, no. 10, 2021.

[5] E. Rslan, M. H. Khafagy, K. Munir, and R. M. Badry, ‘‘English semantic similarity based on map reduce classification for agricultural complaints,’’ Int. J. Adv. Comput. Sci. Appl., vol. 12, no. 12, pp. 1–8, 2021.

[6] W. H. Gomaa and A. A. Fahmy, ‘‘Automatic scoring for answers to Arabic test questions,’’ Comput. Speech Lang., vol. 28, no. 4, pp. 833–857,
Jul. 2014.

[7] N. Y. Habash, ‘‘Introduction to Arabic natural language processing,’’Synth. Lectures Human Lang. Technol., vol. 3, no. 1, pp. 1–187, Jan. 2010.

INTERNATIONAL JOURNAL OF MULTIDISCIPLINARY RESEARCH IN SCIENCE, ENGINEERING AND TECHNOLOGY

| Mobile No: +91-6381907438 | Whatsapp: +91-6381907438 | ijmrset@gmail.com |

www.ijmrset.com