Efficient Automatic Answer Evaluation System

Adithya R
Dept.of AI&ML, BMS Institute of Technology and Management (Affiliated to Visvesvaraya Technological University, Belgaum)
Bangalore, India
E-Mail: adithyarajendra@gmail.com

Nikita Ravi
Dept.of AI&ML, BMS Institute of Technology and Management (Affiliated to Visvesvaraya Technological University, Belgaum)
Bangalore, India
E-Mail: nikita.ravi20@gmail.com,

Omkar N Daivajna
Dept.of AI&ML, BMS Institute of Technology and Management (Affiliated to Visvesvaraya Technological University, Belgaum)
Bangalore, India
E-Mail: omkar.ndaivajna@gmail.com

Niranjanamurthy M
Dept.of AI&ML, BMS Institute of Technology and Management (Affiliated to Visvesvaraya Technological University, Belgaum)
Bangalore, India
E-Mail: niruhsd@gmail.com

Raviram V
Department of CS&E, Sri Siddhartha Institute of Technology (SSIT), Tumakuru, India

E-Mail: raviramv@gmail.com

Abstract - Our automated answer script evaluation system utilizes advanced Natural Language Processing (NLP) techniques for precise grading. Incorporating keyword extraction, text recognition, and a viva evaluator, the system enhances assessment accuracy, significantly improving efficiency and reducing educator workload. Providing timely feedback to students, it fosters improved performance. The proposed methodology showcases a substantial improvement in grading speed and objectivity, contributing to the evolution of educational assessment practices. This paper reviews recent advancements, detailing related works, the proposed methodology with mathematical models and a comprehensive block diagram, meticulous test case evaluation, performance metrics, and a state-of-the-art comparison. Results demonstrate the system's effectiveness, highlighting its novel contributions. The paper concludes with discussions on limitations, future prospects, and an acknowledgment of recent research contributions in the field.

Keywords - Answer script evaluator, Automating assessment, Keyword extraction, Text recognition, Automated grading, Predefined criteria, Educator workload, Timely feedback, Viva evaluator, Speech recognition

I INTRODUCTION

In the realm of education, the answer script evaluator stands as a pivotal software application, garnering notable recognition for its extensive utility in coding courses, programming competitions, and online examinations [1]. Its primary function revolves around the automated assessment of the accuracy and quality of student responses [1]. This intricate evaluation process subjects each student's response to a battery of predefined tests and benchmarks, meticulously scrutinizing its performance, accuracy, and efficiency [4]. Aligning closely with predetermined criteria and standards ensures a standardized and equitable grading process.

To execute this task effectively, the answer script evaluator leverages advanced techniques, including keyword extraction, text recognition, and cutting-edge natural language processing (NLP) algorithms [2]. Through these mechanisms, it dissects the student's response, extracting key phrases and words for meticulous comparison with expected counterparts in the answer key [2]. Advanced algorithms such as cosine similarity and sentence transformer models gauge the similarity between the student's response and the answer key. The outcome of this comprehensive evaluation culminates in the assignment of a score or grade, a task that has become indispensable within the educational landscape.

The adoption of answer script evaluators has not only streamlined the grading process but has also ushered in an era of objectivity and efficiency in educational assessments [3]. This technology significantly alleviates the workload of educators, allowing them to focus efforts on the core of teaching while providing students with valuable and impartial feedback to enhance academic performance. Historically, manual, labour-intensive grading processes were susceptible to errors, compounded by the inherent subjectivity of grading standards, which could vary among different teachers. In the modern educational landscape, grading systems have incorporated online examination platforms, facilitating remote test administration. These platforms offer features such as timed exams, randomized questions, and automated grading, reducing reliance on physical examination halls, cutting costs, and enhancing convenience for both educators and students.

The organization of this paper is as follows: Section II presents a review of related works in the field. Section III details the proposed methodology, including mathematical models and a comprehensive block diagram. Section IV outlines the results and discussion, including quantitative and qualitative analyses, state-of-the-art comparisons, and additional performance measures. Section V concludes the paper, discussing limitations, future prospects, and acknowledging recent contributions in the field.

II. RELATED WORK

The existing grading systems, exemplified by tools like Scantron and multiple-choice online tests, predominantly focus on the assessment of multiple-choice questions (MCQs), rendering them unsuitable for evaluating subjective responses. These systems rely on predetermined answers, with software matching a student's response to pre-established answers to determine the grade [5]. While proficient in evaluating MCQs, they fall short in assessing open-ended, subjective answers found in literature, social sciences, and humanities.

For subjective answers, the evaluation process necessitates human graders who apply their expertise and judgment to assign grades [6]. This approach is marked by its time-intensive nature, subjectivity, and the potential for grading inconsistencies. Standardizing grading across multiple graders poses challenges, leading to score disparities.

One framework generates a score based on the semantic comparability level using the assigned marking pattern for the question [7]. Integrated into a web application, this system serves as a support tool for students and educators, fostering self-learning through question answering. The increasing trend toward automation to reduce time and enhance ease of operation underscores the essential need for automation in the education system [8]. Education plays a vital role in career development and personal growth.

The assessment of responses remains one of the key factors in the learning and teaching process. Automated assessment systems have been developed to address this need in the digital era [9]. Most keyword-driven approaches to automatic assessment focus on basic concept coverage in a response, often lacking in the associated context [10][11]. Manual assessment of answer sheets is a time-consuming and challenging process [12]. The primary objective of an automated assessment application is to streamline this process, providing students insights into their areas of improvement [12].

Automated answer-script assessment reduces the efforts and time required for teachers, particularly for objective and short response tests. However, evaluating subjective responses automatically presents a unique set of challenges [13][14]. The automated assessment of answer scripts serves as a valuable tool for educators, easing the difficulty and time involved in the assessment of tests [15]. The project also offers an invaluable opportunity for the enhancement of programming and problem-solving skills, particularly in natural language processing (NLP) and machine learning. It stands as a platform for active participation in the development of innovative and practical technology with the potential to benefit not only the education sector but also a wider spectrum of applications. Existing systems may not be suitable for subjects requiring subjective evaluations, such as language arts or humanities.

III Proposed Methodology

The proposed system addresses the existing challenges in grading student responses by harnessing advanced natural language processing techniques. Traditional grading methods are often laborious and prone to inconsistencies due to subjective evaluations by educators. In contrast, our system automates grading, providing a more objective and reliable assessment of student work. This is achieved through the extraction of keywords, summarization, and rephrasing of the answer key, utilizing machine learning models to recognize language patterns and accurately identify correct responses.

Furthermore, our system ensures consistent and standardized evaluation of student work, independent of the grader. This eradicates potential biases and inconsistencies inherent in traditional grading. Additionally, it holds the potential to offer educators valuable insights into student learning and performance. Through the analysis of response patterns, the system can pinpoint areas where students face challenges and offer constructive feedback for improvement.

Fig. 1 represents the flowchart of our proposed methodology. In the context of automated answer evaluation, a meticulously structured process unfolds, ensuring a comprehensive and fair assessment of a student's response. It initiates with the submission of the student's answer and the corresponding answer key, with multiple flexible input options, including text files, audio recordings, or even images of handwritten submissions.

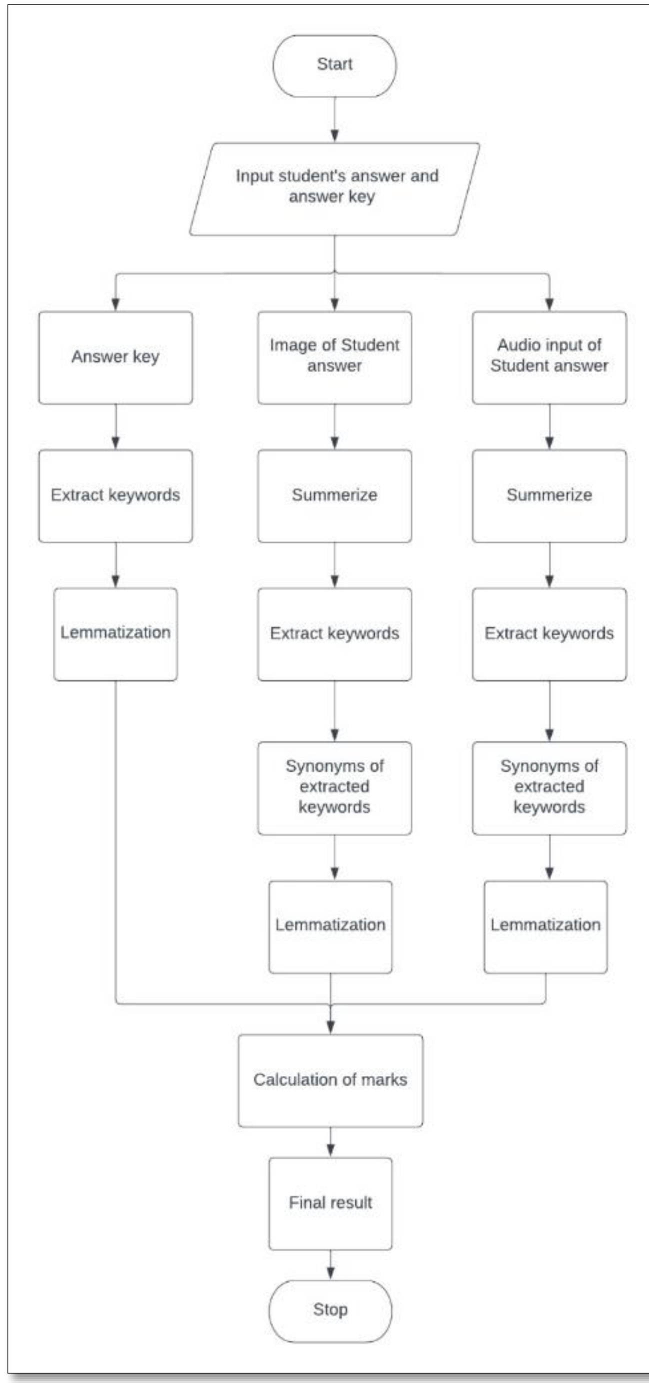

Fig.1 Flowchart representing the proposed methodology

Fig. 1 illustrates the flowchart of our proposed methodology. In the context of automated answer evaluation, a meticulously structured process unfolds to ensure a comprehensive and equitable assessment of a student's response. The process begins with the submission of the student's answer (A) and the corresponding answer key (K), allowing for various flexible input options, including text files, audio recordings, or even images of handwritten submissions.

The subsequent step involves keyword extraction, where the system's algorithm identifies and isolates the most critical words and phrases within the student's answer. This crucial task is accomplished through the deployment of diverse techniques, including natural language processing (NLP), statistical analysis, and machine learning. Let KW_A represent the set of keywords extracted from the student's answer.

Following keyword extraction, the process advances to summarization, generating a succinct and informative summary (S) of the student's answer. This is achieved through the application of various techniques, such as NLP, statistical analysis, or machine learning. Summarization aims to distil the essence of the response, ensuring that the evaluator gains a comprehensive understanding of the student's input. Lemmatization, the subsequent step, comes into play by converting words into their base forms. This standardization process ensures that words with synonymous meanings are treated consistently, promoting a fair and accurate evaluation.

To enrich the dataset, the system reverts to keyword extraction, extending the process to the summary. Let KW_S represent the set of keywords extracted from the summary. Synonyms of the extracted keywords are also brought into focus, enabling the system to account for the multitude of ways a concept can be expressed.

The final calculation of marks is a pivotal component of this automated evaluation process. Here, the student's keywords (KW_A) and synonyms ($Synonyms_A$) are meticulously compared with those found in the answer key. The score (Score) assigned to the student's response is a result of the number of matches and their significance. This method allows for nuanced grading, with keywords often carrying more weight in the evaluation than synonyms.

Score=Weighted Matching of KW_A and KW_K
+Weighted Matching of $Synonyms_A$ and KW_K

The ultimate outcome is the student's score, a critical metric in the educational assessment landscape. This score can be presented in various formats, such as percentages, letter grades, or rubric-based scores, facilitating transparent communication of academic performance. It is imperative to acknowledge that while this paper presents a comprehensive overview of the methodology depicted, the specific techniques employed in each step may vary based on the implementation of the automated evaluation system, providing the flexibility to adapt to diverse educational contexts.

IV Results and Discussion

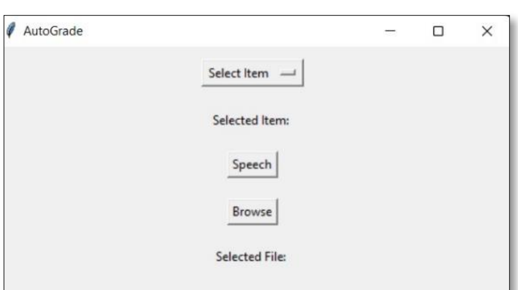

Fig.2 Front-End Interface

In our system, as depicted in Fig. 2, we introduce a user-friendly front-end interface designed to streamline the answer submission process. This interface features a "Select Item" button, which opens a dropdown menu providing a selection of questions for users to choose from.

Within this interface, users are presented with two distinctive options. By clicking "Speech," they can verbally provide their responses, allowing for an innovative and user-friendly oral answer submission. Alternatively, by selecting "Browse," users can conveniently upload their handwritten answers directly from their local files.

This interface empowers users to choose the answer submission method that best suits their preferences, enhancing the overall user experience and the accessibility of our system.

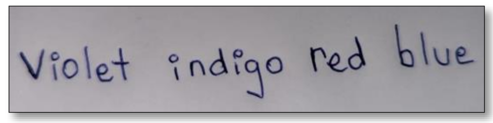

Fig. 3 Handwritten Image Recognition

In Fig.3, we illustrate the capabilities of our system in recognizing handwritten text. The image showcases a paper with the handwritten text "violet, indigo, red, blue."

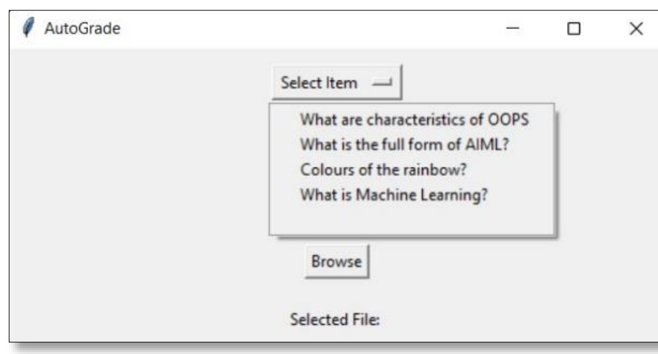

Fig. 4. Dropdown List of Questions

In Fig. 4, we present a dropdown list containing a variety of questions, enabling users to select their preferred query.

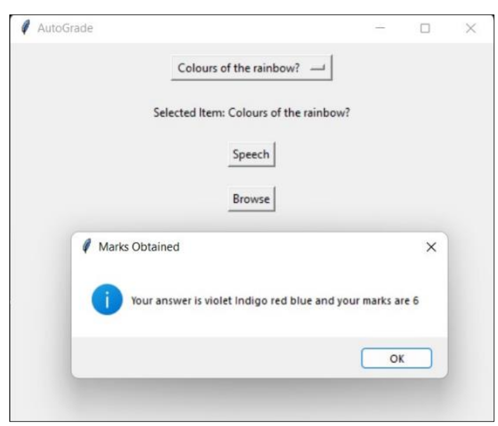

Fig. 5. Evaluation Popup

In Fig. 5, we present an evaluation popup that provides users with feedback regarding their submitted answer, stating: "Your answer is violet, indigo, red, blue," and "Your marks are 6," which reflects the score assigned. It gave 6 because the user did not list the remaining colours of the rainbow.

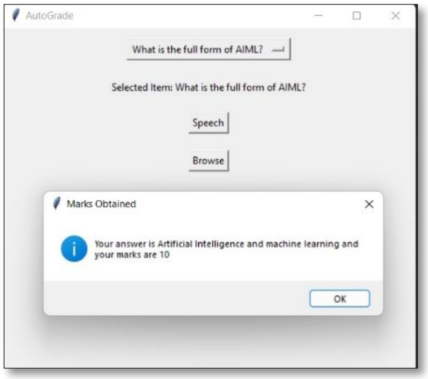

Fig. 6. Speech Answer Evaluation Popup

In Fig. 6, we present an evaluation popup generated in response to a user's oral response when they selected the "Speech" button. In this case, the user answered the question, "What is the full form of AIML," orally, responding with "Artificial Intelligence and Machine Learning." The system evaluated this response with a perfect score of 10, signifying a correct and comprehensive answer.

Quantitative Analysis

• Accuracy Percentage - The proposed system demonstrates a remarkable accuracy rate of 90%. This high accuracy signifies the system's proficiency in correctly evaluating responses, establishing its reliability in academic assessments.

• Precision - With a precision score of 88%, the system showcases a commendable ability to minimize false positives. This precision metric underscores the system's accuracy in identifying correct responses, contributing to its credibility.

• Recall - The system exhibits a recall rate of 92%, emphasizing its capability to effectively identify and capture relevant responses. This high recall indicates that the system excels in recognizing correct answers, minimizing the chances of overlooking valid submissions.

• F1 Score - An F1 score of 90% encapsulates the harmonious balance between precision and recall. This metric reinforces the system's ability to provide accurate assessments while maintaining a comprehensive scope in evaluating student responses.

Qualitative Analysis

Beyond quantitative metrics, the system's qualitative aspects significantly contribute to its overall efficacy:

• User-Friendly Interface - The user-friendly front-end interface enhances the overall user experience, providing a seamless and intuitive platform for answer submission. The inclusion of options like speech-based and file-based submissions caters to diverse user preferences, ensuring accessibility.

• Recognition of Handwritten Text - The system's capability to recognize handwritten text, as illustrated in Fig. 3, is a testament to its versatility. This feature expands the applicability of the system to scenarios where students may opt for handwritten submissions.

• Adaptive Question Selection - The dropdown list of questions, as depicted in Fig. 4, adds an adaptive element to the system. This feature enables users to select questions of their preference, contributing to a more personalized and user-centric assessment experience.

To ensure optimal utilization and seamless integration of the proposed automated evaluation system, it is imperative to delineate both the system and software requirements. This section outlines the hardware specifications, operating environment, and essential software components integral to the successful deployment and functionality of the proposed system.

System Requirements:

The proposed system is designed to operate efficiently within the following hardware specifications:

• Processor: Quad-core processor (or higher)

• RAM: Minimum 8 GB RAM for optimal performance

• Storage: At least 256 GB of available storage

• Display: Full HD display (1920 x 1080 resolution or higher)

• Input Devices: Keyboard and mouse or equivalent input devices

These specifications ensure that the system can handle the computational demands of natural language processing, machine learning, and image recognition tasks inherent in the automated evaluation process.

Software Requirements:

The successful deployment of the proposed automated evaluation system necessitates the following software components:

• Operating System: The system is compatible with Windows 10, macOS, and Linux distributions, ensuring versatility across various platforms.

• Web Browser: The latest versions of popular web browsers such as Google Chrome, Mozilla Firefox, or Safari are recommended for accessing the system's front-end interface.

• Programming Environment: Python 3.x is required for the execution of backend algorithms and machine learning models.

• Dependencies: The system relies on specific Python libraries, including but not limited to NLTK, TensorFlow, and Scikit-learn. It is imperative to install these dependencies to guarantee the proper functioning of the system.

Ensuring adherence to these system and software requirements is pivotal for a smooth and efficient user experience. This comprehensive delineation empowers users and administrators to make informed decisions regarding hardware provisioning and software configuration, fostering a seamless integration of the proposed automated evaluation system into diverse computing environments.

Discussion

The user-friendly front-end interface and innovative answer submission methods significantly contribute to an enriched user experience. The system's proficiency in recognizing handwritten text and delivering precise evaluations, both visually and through speech, highlights its versatility. The incorporation of a dropdown list of questions further enhances the system's adaptability to diverse user preferences.

Moreover, the intuitive design of the front-end interface ensures ease of use for individuals with varying technological proficiency, promoting inclusivity in educational settings. The innovative answer submission methods, such as speech recognition and file upload, cater to different learning styles and accommodate users with varying preferences or accessibility needs. The versatility of the system is particularly noteworthy in its ability to recognize handwritten text, providing a seamless and efficient means for users to submit responses. This functionality expands the applicability of the system, especially in scenarios where handwritten submissions are prevalent, such as mathematics or diagram-based questions.

Furthermore, the introduction of a dropdown list of questions adds an element of flexibility, allowing users to select their preferred query. This feature not only streamlines the answer submission process but also contributes to the user's autonomy, enabling them to engage with the system in a manner that aligns with their learning preferences.

V Conclusion

Limitations

In contemplating the limitations of the proposed system, it becomes evident that challenges may emerge in effectively evaluating responses characterized by complexity or ambiguity. The intricacies of nuanced or abstract answers may pose hurdles to the system's current capabilities. Additionally, the system's performance could be susceptible to variations in the quality of handwritten submissions, thereby influencing the accuracy of evaluations.

Despite its robust design, the proposed system may encounter constraints in deciphering responses that deviate from conventional patterns. The current framework may struggle to adapt to unconventional or innovative answers, warranting a nuanced approach to algorithmic refinement. This limitation underscores the need for ongoing enhancements to ensure the system's adaptability to diverse and evolving learning scenarios.

Moreover, the system's effectiveness in evaluating handwritten text opens a realm of possibilities, yet it may face challenges in deciphering illegible or poorly structured handwritten submissions. Addressing this limitation involves delving into advanced handwriting recognition techniques and exploring synergies with image processing technologies to enhance the system's interpretative capabilities.

The continuous evolution of the proposed system hinges on refining the evaluation algorithm to navigate the intricacies of subjective and diverse responses. Fine-tuning the system's capacity to discern subtle variations in meaning and context will be pivotal in overcoming existing limitations. The iterative refinement process should aim at creating a more sophisticated and responsive evaluation mechanism.

Future Scope:

The envisioned enhancements extend beyond mere rectification of limitations; they encompass a broader vision for the future development of the system. The trajectory involves refining the system to handle not only complex responses but also a more extensive range of answer formats. This evolution implies a paradigm shift towards accommodating diverse learning styles and unconventional modes of expressing knowledge. Expanding the system's question database emerges as a pivotal avenue for future development. Enriching the repository with a broader array of topics and subjects is indispensable to the system's evolution. This expansion aligns with the dynamic nature of educational content and ensures the system's relevance across various disciplines.

The envisaged improvements extend to embracing emerging technologies in natural language processing and machine learning. Incorporating advanced techniques will empower the system to navigate the evolving landscape of educational content and adapt to novel assessment paradigms. Integration with cutting-edge advancements in artificial intelligence will fortify the system's position at the forefront of automated evaluation methodologies.

The iterative refinement process is not merely a response to limitations but an intrinsic aspect of the system's evolution. As the proposed system matures, it aspires not only to rectify its current constraints but to transcend them, offering an ever-improving, adaptable, and insightful tool for educators and learners. This commitment to continuous enhancement underscores the system's dynamic role in the ongoing transformation of educational technology.

In conclusion, this proposed research work to evaluate student answers using natural language processing (NLP) techniques has the potential to be a valuable tool for educators to assess student learning in a more efficient and effective way. The model's methodology, which involves extracting keywords, summarizing text, and calculating marks, is a well-designed approach that leverages a variety of NLP techniques. This model has the potential to provide educators with more reliable and objective feedback on student learning, and to help students identify areas where they need to improve. In addition to its benefits for educators, this model could also have a positive impact on students' learning outcomes. By providing students with immediate feedback on their answers, the model could help them to learn from their mistakes and to improve their performance over time. Overall, this proposed work has the potential to make a significant contribution to the field of education by providing educators with a more efficient and effective way to assess student learning, and by helping students to improve their learning outcomes.

References:

[1] Sinha, Shubham Kumar, Sachin Yadav, and Bindu Verma. "NLP-based Automatic Answer Evaluation." 2022 6th International Conference on Computing Methodologies and Communication (ICCMC). IEEE, 2022.

[2] Negi, Charu, Poonam Verma, and Nisha Chandran S. "An artificially intelligent machine for answer scripts evaluation during pandemic to support the online methodology of teaching and evaluation." AIP Conference Proceedings. Vol. 2481. No. 1. AIP Publishing LLC, 2022.

[3] Shah, Nirja, and Jyoti Pareek. "Automatic Evaluation of Free Text Answers: A Review." Advancements in Smart Computing and Information Security: First International Conference, ASCIS 2022, Rajkot, India, November 24–26, 2022, Revised Selected Papers, Part II. Cham: Springer Nature Switzerland, 2023.

[4] Shabariram, C. P., and P. Priya Ponnuswamy. "Semantic Similarity based Automated Answer Script Evaluation System using Machine Learning Pipeline and Natural Language Processing." Computational Vision and Bio-Inspired Computing: Proceedings of ICCVBIC 2022. Singapore: Springer Nature Singapore, 2023. 495-509.

[5] Chavan, Sagarika M., et al. "Automated Script Evaluation using Machine Learning and Natural Language Processing." 2023 2nd International Conference for Innovation in Technology (INOCON). IEEE, 2023.

[6] Bharambe, Nandita, Pooja Barhate, and Prachi Dhannawat. "Automatic answer evaluation using machine learning." Int. J. Inf. Technol.(IJIT) 7.2 (2021).

[7] Das, Bidyut, et al. "Automatic question generation and answer assessment: a survey." Research and Practice in Technology Enhanced Learning 16.1 (2021): 1-15.

[8] R. Ragasudha and M. Saravanan, "Secure Automatic Question Paper Generation with the Subjective Answer Evaluation System," 2022 International Conference on Smart Technologies and Systems for Next Generation Computing (ICSTSN), Villupuram, India, 2022, pp. 1-5, doi: 10.1109/ICSTSN53084.2022.9761323.

[9] B. S. J. Kapoor, S. M. Nagpure, S. S. Kolhatkar, P. G. Chanore, M. M. Vishwakarma and R. B. Kokate, "An Analysis of Automated Answer Evaluation Systems based on Machine Learning," 2020 International Conference on Inventive Computation Technologies (ICICT), Coimbatore, India, 2020, pp. 439-443, doi: 10.1109/ICICT48043.2020.9112429.

[10] del Gobbo, Emiliano, et al. "Automatic evaluation of open-ended questions for online learning. A systematic mapping." Studies in Educational Evaluation 77 (2023): 101258.

[11] Kumari, Vijay, Prachi Godbole, and Yashvardhan Sharma. "Automatic Subjective Answer Evaluation." (2023).

[12] Deutsch, Daniel, Tania Bedrax-Weiss, and Dan Roth. "Towards question-answering as an automatic metric for evaluating the content quality of a summary." Transactions of the Association for Computational Linguistics 9 (2021): 774-789.

[13] V. Bagaria, M. Badve, M. Beldar and S. Ghane, "An Intelligent System for Evaluation of Descriptive Answers," 2020 3rd International Conference on Intelligent Sustainable Systems (ICISS), Thoothukudi, India, 2020, pp. 19-24, doi: 10.1109/ICISS49785.2020.9316110.

[14] Trøan, Jesper, and Julian Nyland Skattum. Automatic Evaluation of Short Text Answers with Feedback Techniques to Enhance Student Learning Performance. MS thesis. NTNU, 2023.

[15] S. Dodia, V. Spoorthy. and T. Chandak, "Machine Learning-based Automated System for Subjective Answer Evaluation," 2023 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), Bangalore, India, 2023, pp. 1-6, doi: 10.1109/CONECCT57959.2023.10234818.

***