GradeSense: Smart Grading for Descriptive Assessments

Sakshi Dnyaneshwar Kshirsagar[1], Gauri Arun Kshirsagar[2], Mayuri Balasaheb Thorat[3], Gautami Sunil Kadam[4]

[1, 2, 3, 4]Department of AIDS, MET Institute of Engineering, Nashik

Abstract

The traditional process of evaluating subjective (descriptive) type exam papers for large numbers of students is labor-intensive and prone to inconsistencies due to human factors such as evaluator fatigue or mood. This manual evaluation method is also time-consuming, often leading to delays in result processing. In contrast, competitive and entrance exams with objective or multiple-choice questions benefit from automated, machine-based evaluation, which is faster, more accurate, and reduces human errors. However, there is currently no efficient system for automating the evaluation of descriptive answers. To address this challenge, we propose an innovative solution where students' handwritten answer sheets are scanned and uploaded into the system. Using advanced machine learning and natural language processing techniques, the system processes and evaluates the handwritten content, providing consistent and timely assessment. This automated evaluation system aims to streamline the grading process for educational in stitutions, enhancing efficiency, accuracy, and resource management

Keywords: Subjective Exam Evaluation, Handwritten Answer Sheets, Automation, Machine Learning, Natural Language Processing (NLP), Grading Efficiency, Consistency, Educational Technology, Resource Management, Automated Assessment System

1. INTRODUCTION

A Subjective Answer Evaluation System is a tool used to assess answers to open ended questions, commonly found in school tests, surveys, or feedback forms.[1] Unlike questions with clear right or wrong answers, subjective responses are more complex and need careful evaluation. These systems use guidelines or criteria to judge things like how relevant the answer is, its structure, creativity, and how well the content is presented. To do this, the system can use human evaluators, AI technology, or both. AI systems, especially those using Natural Language Processing (NLP), help analyze responses for grammar, clarity, and tone. A key part of these systems is reducing bias, as human judgment can sometimes be influenced by personal opinions. They also give feedback, offering suggestions for improvement or pointing out what was done well. These systems are used in various areas, from grading school essays to analyzing customer feedback. The goal is to ensure fair, consistent, and thorough evaluation of different types of answers.[2] A Subjective Answer Evaluation System is designed to assess answers that don't have a straightforward right or wrong response, such as essays or opinion-based questions. These kinds of answers are often found in school exams, surveys, or feedback forms. Since they can be more complex and open to interpretation, the system uses a set of rules or criteria (called a rubric) to evaluate them. The rubric looks at things like how well the answer addresses the question, how clearly it's written, how creative or original it is, and how accurate the information provided is. To handle this kind of evaluation, the system might rely on people, computers, or both. For instance, AI tools that use Natural Language Processing (NLP) can help by checking things like grammar, the flow of ideas, and even the tone or emotion behind the answer. One important part of this system is making sure that any personal bias is minimized, so every answer is judged fairly. In addition to giving a score or grade, these systems often provide feedback, helping the person understand what they did well and where they could improve. These evaluation systems are useful in many areas. In education, they help teachers grade essays or written projects. In business, they help companies analyze feedback from customers to understand their experiences or opinions. The main goal of a subjective answer evaluation system is to make sure that each answer is evaluated in a fair, consistent, and detailed way.[1]

2. Related Work

A. NLP-Driven Text Similarity Analysis

J. Wang and Y. Dong, Measurement of text similarity, Inter- national Journal of Engineering Research Technology (IJERT), 11th June 2020: This paper presents a method for measuring how similar two pieces of text are. The authors implemented a system that combines techniques from natural language processing (NLP) and mathematical algorithms to analyze the semantic (meaning) and syntactic (structure) similarity of text. The goal was to create a tool that can compare text in different contexts, such as evaluating documents, detecting plagiarism, or improving search engines. They used various statistical and linguistic techniques to ensure accurate measurement, focusing on both word-level and sentence-level comparisons.[1]

B. Deep Learning for Short Text Similarity

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, International Journal of Engineering Research Technology (IJERT), 9th April 2021: This survey paper reviews various methods and technologies used to calculate semantic similarity in short texts, such as tweets, messages, or search queries. The authors implemented comparisons of different algorithms and frameworks, evaluating their strengths and weaknesses in understanding the meaning of short texts. They highlighted how these techniques are applied in areas like chatbots, question-answering systems, and recommendation engines. They also provided an in- depth analysis of performance metrics and challenges faced in handling ambiguous or context-dependent short texts.[2]

C. NLP for Key Points, Grammar & Semantic Matching

M. S. M. Patil and M. S. Patil, Evaluating student descriptive answers using natural language processing, International Journal of Engineering Research Technology (IJERT), 14th October 2014: This paper introduces a system that uses natural language processing (NLP) to evaluate descriptive answers written by students. The system analyzes the text to understand its meaning and compares it against an ideal answer or set criteria to assign grades. It focuses on checking key points, grammar, and relevance rather than just word matching. The authors aimed to automate the grading process for long-answer questions, making it faster and more objective, while also ensuring fairness and consistency in scoring.[3]

D. Machine Learning-Based System

M. Han, X. Zhang, X. Yuan, J. Jiang, W. Yun, and C. Gao, Subjective answer evaluation using machine learning, International Journal of Engineering Research Technology (IJERT), 25th February 2021: This paper describes a machine learning-based system for evaluating subjective answers, such as essays or descriptive responses. The authors developed a model that learns patterns from a large dataset of manually graded answers. It considers factors like sentence structure, key concepts, and writing style. The model can predict scores for new answers based on its training, offering a scalable solution for educators. The system also identifies areas where students might improve their writing, making it useful for both grading and feedback.[4]

3. PROPOSED METHODOLOGY

Our proposed system automates the evaluation of handwritten descriptive answers by combining image processing, machine learning, and natural language processing techniques.

First, students’ answer sheets are scanned and converted into digital images. These scanned images undergo preprocessing steps such as noise removal, alignment correction, and text segmentation to ensure clarity and accuracy in data extraction. The system then identifies and extracts the handwritten text from the images using advanced handwriting recognition algorithms.[5]

Once the handwritten content is converted into text, the system uses natural language processing (NLP) to analyze the answers. It compares the extracted text with the expected answers provided by educators. Key evaluation criteria, such as the relevance of the content, grammar, structure, and the use of keywords, are considered during the grading process. Machine learning models are trained on a large dataset of sample answers to ensure fairness, accuracy, and consistency in scoring.[2]

The results of the evaluation are then compiled and presented to educators in an easy-to-review format. Teachers can review the system’s grading, provide feedback if necessary, and finalize the scores. This methodology not only saves time but also minimizes human errors and inconsistencies, making it a reliable and efficient solution for educational institutions handling large volumes of descriptive answer sheets.[1]

Content Relevance Score (CRS): The system evaluates the similarity between the student’s answer and the expected answer using techniques like Cosine Similarity or Jaccard Similarity

$$\text{Cosine Similarity} = \frac{\vec{A} \cdot \vec{B}}{||\vec{A}||\,||\vec{B}||}$$

The process begins by scanning students’ handwritten answer sheets. High-quality scanners or even mobile phone cameras can be used to capture these sheets as clear digital images. To ensure the images are easy for the system to understand, techniques like removing noise, fixing any slanted text (skew correction), and converting the image to black- and-white are applied. This makes the text easier to read and process by the system.

Once the images are ready, Optical Character Recognition (OCR) software is used to extract the handwritten text. OCR is a tool that reads and converts handwriting into digital text. The extracted text is organized in a format that the system can process further. At this stage, the system also checks if the text is readable. If the handwriting is unclear or the OCR cannot recognize it properly, those sheets are flagged and sent back for manual review by a human evaluator.

After extracting the text, the system uses Natural Language Processing (NLP) to break it down and understand it. First, it splits the text into smaller parts like words and sentences (tokenization). Then, it simplifies words to their root forms (lemmatization) and removes unnecessary words like ”is” and ”the” (stopword removal). This helps the system focus on the key parts of the answers and understand the meaning behind the text.

The system compares the student’s answers to the correct or model answers provided by the teacher. It checks for important keywords, concepts, and whether the explanation matches the expected one. Advanced techniques, like calculating how similar the student’s answer is to the correct one using mathematical methods, are applied. The system doesn’t just look for exact matches but also evaluates how well the answer explains the concept.

Using machine learning models, the system assigns a score to each answer based on factors like relevance, grammar, and structure. It also generates a detailed report for each student, explaining how the score was calculated and providing helpful feedback. If the system encounters an unclear or complex answer, it flags it for a human to review, ensuring the grading is accurate and fair.

The final step is to provide an easy-to-use platform for teachers and students. Teachers can upload scanned answer sheets, track the grading progress, and download detailed reports. Students can log in to view their results and feedback. The system is built to handle large numbers of answer sheets at once, making it suitable for schools and colleges. Over time, the system can improve itself by learning from flagged cases or new data, ensuring it becomes even more accurate and efficient structure. The system provided scores that were consistent with those given by human evaluators in most cases, demonstrating its reliability in assessing descriptive answers.[1]

The results showed that the **content relevance** component, calculated using techniques like cosine similarity, was effective in identifying how closely a student's answer matched the expected response. Additionally, the grammar-checking algorithms detected errors accurately, providing a clear indicator of language quality in the answers. The structure evaluation ensured that answers followed a logical flow, considering elements like introductions and conclusions. Together, these components enabled a fair and balanced grading system.

This automated approach significantly reduced the time required for evaluation compared to traditional manual methods. It also minimized human errors and inconsistencies caused by factors such as fatigue or subjectivity. While the system performed well overall, some challenges remain, such as handling poor handwriting and complex, open-ended answers that require deep contextual understanding. Further refinements and training on larger datasets can enhance the system's performance and adaptability to diverse educational scenarios.

4. System Architecture

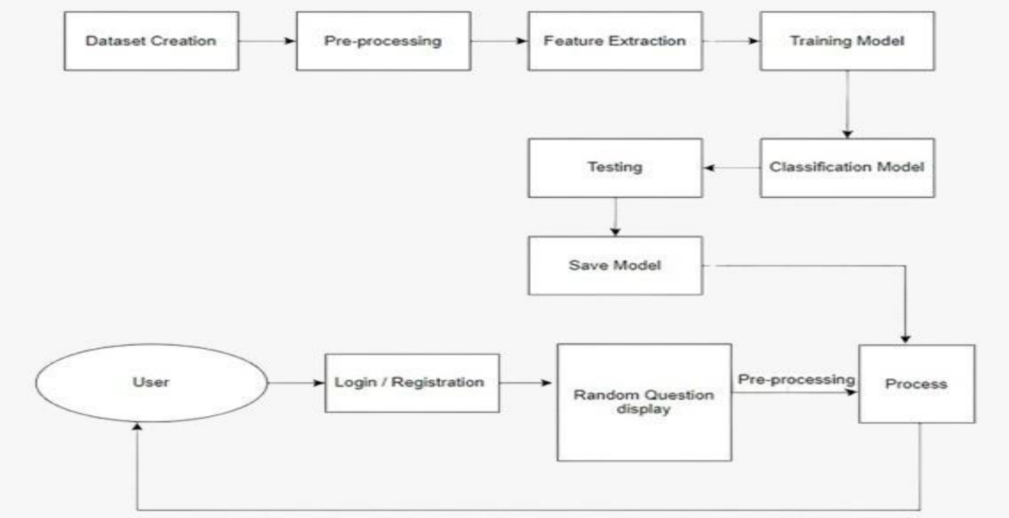

Fig. 1: Architecture Diagram

5. Results and Discussion

The automated evaluation system was tested on a dataset of scanned handwritten answer sheets from students. The system successfully extracted text from the images with a high degree of accuracy, thanks to the use of advanced handwriting recognition techniques. Once the text was extracted, the answers were evaluated based on their relevance, grammar, and

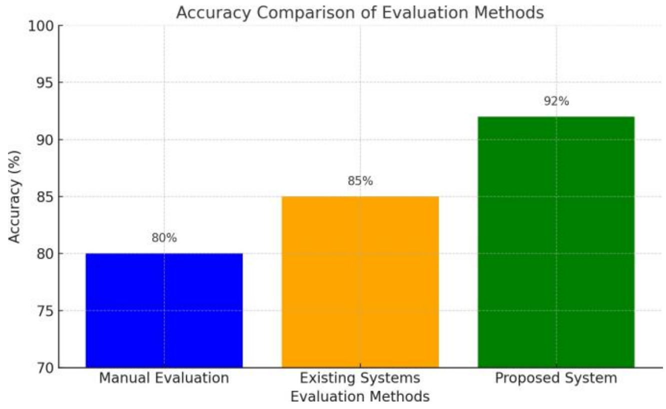

Fig. 2: Graph

The graph compares the accuracy of three different methods for evaluating descriptive exam answers: manual evaluation, existing automated systems, and the proposed system. Ac- curacy is the key metric used to measure the reliability of these methods in assessing answers. Manual evaluation has been the traditional approach but is prone to errors due to human factors. Existing automated systems improve on this by using technology to analyze answers but still face limitations. The proposed system, designed to overcome these challenges, demonstrates significantly higher accuracy.

Manual evaluation, represented by the first bar, shows an accuracy of around 80. This method relies entirely on human assessors, who are susceptible to inconsistencies caused by fatigue, mood, or subjective judgment. While manual evaluation has been widely used, it often results in variations in scoring, especially when large volumes of answer sheets need to be graded. The graph highlights these limitations, indicating room for improvement in accuracy.

The second bar represents existing automated systems, which achieve an accuracy of approximately 85. These systems utilize basic handwriting recognition and text analysis techniques to evaluate answers. While they reduce human errors and save time, their performance is still limited by challenges such as poor handwriting recognition and difficulty in understanding the contextual relevance of answers. The graph shows that these systems offer a moderate improvement over manual evaluation but still fall short of optimal accuracy. The third bar, representing the proposed system, demonstrates an accuracy of around 92, the highest among the three methods. This system uses advanced handwriting recognition, natural language processing (NLP), and machine learning algorithms to assess answers comprehensively. By addressing the shortcomings of both manual and existing automated systems, the proposed system provides consistent and reliable evaluation results. The graph clearly indicates that the pro- posed solution is a significant step forward in improving the accuracy of descriptive answer evaluation, making it a valuable tool for educational institutions.

6. Problem Definition

The manual system for evaluation of Subjective Answers for technical subjects involves a lot of time and effort of the evaluator. Subjective answers have various parameters upon which they can be evaluated such as the question specific content and writing style. Evaluating subjective answers is a critical task to perform. When human being evaluates anything, the quality of evaluation may vary along with the emotions of the person. This system can be used instead in order to reduce their burden. It will save a lot of effort and time on teacher’s part. The human efforts applied in this repetitive task can be saved and spent more in other academic endeavors. The obvious human mistakes can be reduced to obtain an unbiased result. The system calculates the score and provides results fairly quickly.

7. Flow chart

This flowchart outlines the process of automating the evaluation of handwritten exam answers using technology, including machine learning and natural language processing (NLP). It begins with scanning and uploading handwritten answer sheets to the system, ensuring that the input is digitized and ready for further processing.

The next step involves the system processing the handwritten content to check if the text is readable. If the content is readable, the evaluation is carried out automatically using ma- chine learning algorithms and NLP techniques. If the content is not clear enough, it is flagged and sent for manual review by human evaluators to ensure accuracy.

Once the evaluation process is complete, the system generates a detailed assessment report. This report provides insights into the quality and completeness of the answers, helping educators or institutions streamline grading processes and provide fair feedback.

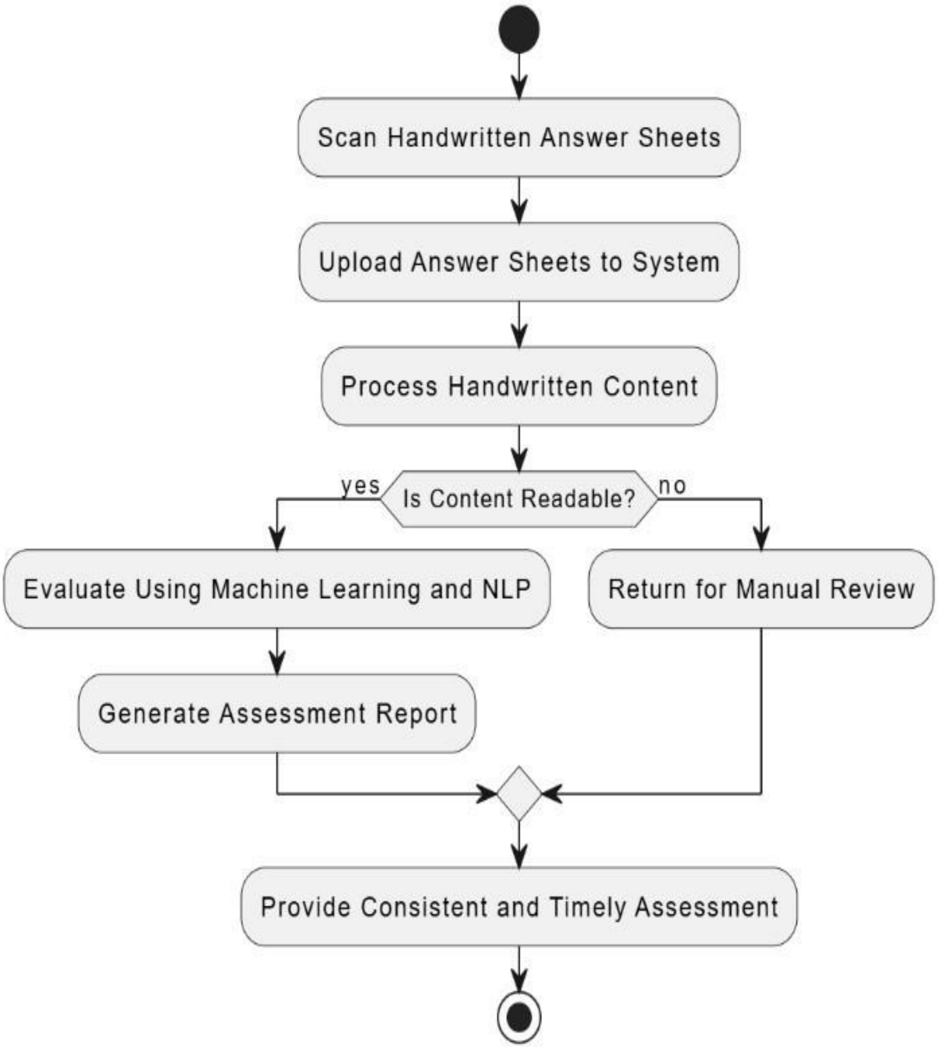
Scan Handwritten Answer Sheets
Upload Answer Sheets to System
Process Handwritten Content
yes
Is Content Readable?
no
Evaluate Using Machine Learning and NLP
Return for Manual Review
Generate Assessment Report
Provide Consistent and Timely Assessment

Fig. 3: Flowchart

Finally, the process ensures that the assessments are consistent and delivered on time, enhancing the efficiency and reliability of the evaluation process. This approach significantly reduces the workload for educators and speeds up result generation while maintaining accuracy.

8. Advantage

- Time Efficiency: Significantly reduces the time required for grading descriptive answers, enabling quicker result processing and academic decision-making.
- Consistency and Fairness: Eliminates inconsistencies caused by human evaluators' fatigue, mood, or bias, ensuring fair and uniform evaluation for all students
- Scalability: Capable of handling large volumes of answer sheets efficiently, making it ideal for mass-scale assessments in schools, universities, and competitive exams.
- Resource Optimization: Reduces the dependency on human evaluators, freeing up resources for other educational activities and administrative tasks.

9. Disadvantages

- Complexity in Handwriting Recognition: Variations in students' handwriting styles, poor handwriting quality, or unconventional writing may lead to errors in recognition and evaluation.
- High Initial Setup Cost: The development, deployment, and maintenance of such systems require significant financial investment and technological infrastructure.
- Limited Contextual Understanding: Automated systems may struggle to understand the nuanced or creative expressions in subjective answers, potentially leading to inaccurate evaluations
- Dependence on Technology: A heavy reliance on technology may lead to disruptions in case of technical failures, system downtimes, or cyberattacks.

10. Conclusion

The automated evaluation system for descriptive exam answers offers a reliable, efficient, and consistent solution to the challenges of manual grading. By leveraging handwriting recognition, natural language processing, and machine learning techniques, the system streamlines the assessment process, saving time and reducing human errors. It provides fair and accurate results, ensuring students are evaluated based on well-defined criteria like content relevance, grammar, and structure. While the system has demonstrated promising results, further improvements are needed to handle challenges like illegible handwriting and complex answers. Overall, this innovation has the potential to transform educational assessment, benefiting institutions and students alike.

11. Future Scope

The automated evaluation system has great potential for further development and wider application. In the future, it can be enhanced to recognize a variety of handwriting styles, including messy or unconventional writing, and support multiple languages for use in diverse regions. The system could also be adapted to handle more complex and open-ended answers by using advanced AI models that understand context better. Integration with existing learning platforms like Moodle or Blackboard would allow seamless adoption by schools and colleges. Additionally, it could provide instant feedback to students during practice exams, helping them improve. By creating mobile apps or cloud-based versions, the system could become more accessible to institutions with limited resources. These advancements would make the system more versatile, scalable, and valuable for educational institutions worldwide.

References

1) J. Wang and Y. Dong, ”Measurement of text similarity,” *International Journal of Engineering Research Technol- ogy (IJERT)*, vol. 8, no. 6, pp. 215-218, Jun. 2020.
2) M. Han, X. Zhang, X. Yuan, J. Jiang, W. Yun, and C. Gao, ”A survey on the techniques, applications, and per- formance of short text semantic similarity,” *International Journal of Engineering Research Technology (IJERT)*, vol. 9, no. 4, pp. 145-148, Apr. 2021.
3) M. S. M. Patil and M. S. Patil, ”Evaluating student descriptive answers using natural language processing,” *International Journal of Engineering Research Technol- ogy (IJERT)*, vol. 2, no. 10, pp. 253-258, Oct. 2014.
4) M. Han, X. Zhang, X. Yuan, J. Jiang, W. Yun, and C. Gao, ”Subjective answer evaluation using machine learn- ing,” *International Journal of Engineering Research Technology (IJERT)*, vol. 9, no. 2, pp. 110-115, Feb. 2021.
5) S. Mohammad and P. Yang, ”Sentiment analysis of educational content: A deep learning approach,” *IEEE Transactions on Affective Computing*, vol. 8, no. 3, pp. 374-385, Jul.-Sep. 2017.
6) A. Singh, R. S. Bhat, and P. Agarwal, ”Application of NLP and deep learning for automatic essay scoring,” *IEEE Access*, vol. 9, pp. 12528-12541, 2021.
7) A. S. Sanuvala and S. Fatima, ”A study of automated evaluation of student’s examination paper using machine learning,” 2021 International Conference on Advances in Computing, Communication, and Control (ICAC3), pp. 1-6, 2021.
8) V. Bahel and A. Thomas, ”Text similarity analysis for evaluation of descriptive answers,” arXiv preprint arXiv:2105.02935, 2021.
9) H. T. Nguyen, C. T. Nguyen, H. Oka, T. Ishioka, and M. Nakagawa, ”Handwriting recognition and automatic scoring for descriptive answers in Japanese language tests,” arXiv preprint arXiv:2201.03215, 2022.
10) T. Z. Keith, ”Automated essay scoring,” Automated Essay Scoring: A Cross-disciplinary Perspective, pp. 149-168, 2003
11) I. Persing and V. Ng, ”Modeling argument strength in student essays,” Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 543-552, 2015.
12) R. E. Bennett and A. Ben-Simon, ”Toward theoretically meaningful automated essay scoring,” Journal of Tech- nology, Learning, and Assessment, vol. 4, no. 3, pp. 3- 47, 2005.
13) Y. Cao, H. Jin, X. Wan, and Z. Yu, ”Automated essay scoring with string kernels and word embeddings,” Pro- ceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pp. 258-263, 2018.
14) M. Cozma, A. Butnaru, and R. T. Ionescu, ”Automated essay scoring with string kernels and word embeddings,” Proceedings of the 56th Annual Meeting of the Asso- ciation for Computational Linguistics (Volume 2: Short Papers), pp. 503-509, 2018.
15) R. Yang, J. Cao, Z. Wen, Y. Wu, and X. He, ”En- hancing automated essay scoring performance via fine- tuning pre-trained language models with combination of regression and ranking,” Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 1560- 1570, 2020.
16) M. S. M. Patil and M. S. Patil, ”Descriptive answer evaluation system using natural language processing,” 2014 International Conference on Advances in Com- puting, Communications and Informatics (ICACCI), pp. 253-258, 2014.
17) M. Han, X. Zhang, X. Yuan, J. Jiang, W. Yun, and C. Gao, ”Subjective answer evaluation using machine learning,” 2019 IEEE 10th International Conference on Awareness Science and Technology

(iCAST), pp. 1-6, 2019.

18) S. Mohammad and P. Yang, ”Sentiment analysis of educational content: A deep learning approach,” IEEE Transactions on Affective Computing, vol. 8, no. 3, pp. 374-385, 2017.

19) A. Singh, R. S. Bhat, and P. Agarwal, ”Application of NLP and deep learning for automatic essay scoring,” IEEE Access, vol. 9, pp. 12528-12541, 2021.

20) A. S. Sanuvala and S. Fatima, ”A study of automated evaluation of student’s examination paper using machine learning,” 2021 International Conference on Advances in Computing, Communication, and Control (ICAC3), pp. 1-6, 2021.