NLP and OCR based Automatic Answer Script Evaluation System

Pranav Deepak
Dept. of ISE
BMS College of Engineering
Bangalore, India

R. Rohan
Dept. of ISE
BMS College of Engineering
Bangalore, India

R. Rohith
Dept. of ISE
BMS College of Engineering
Bangalore, India

Roopa R., PhD
Dept. of ISE
BMS College of Engineering
Bangalore, India

ABSTRACT
Evaluation of answer scripts is a tedious laborious process in the education domain. Proper solution is proposed in this paper using two different state of art technologies i.e., Natural Language Processing (NLP) and Optical Character Recognition (OCR). To develop the Automatic Answer Script Evaluation System. The system is intended to simplify grading by automatic the scoring of written responses in a consistent and accurate way. The NLP portion of the system is responsible for understanding the semantic purposes in textual content of answer scripts. It uses state-of-the-art language models to assess and infer the context, coherence, and entailment properties of the generated text answers. Using NLP to understand text, the system can check not only for correct grammar but also gauge how deeply a particular concept is understood.

Keywords
Natural Language Processing, OCR Analysis

1. INTRODUCTION
In the field of education, evaluation is a very important mechanism used to measure students understanding as well in terms of mastery over various subjects. Traditionally, this has taken a form of manual and time-consuming process which is also subject to certain degree subjective player biases. For that reason, a push towards an automated solution! But as technology tools like Natural Language Processing (NLP), Optical Character Recognition (OCR), etc, have started evolving and ramping up that are equipped to automate the evaluation process drastically. As the current shift changed everything, implementation of Automatic Answer Script Evaluation Systems (AASES) has been initiated, wherein through NLP and OCR techniques; these systems analyze and evaluative student’s responses more effectively in an impartial way.

NLP is a subfield of AI that enables computers to understand, interpret and generate human language in ways that are difficult for ordinary humans to read. By leveraging techniques such as machine learning, deep learning, and statistical modelling, NLP systems can process vast amounts of textual data and extract valuable insights. In the context of AASES, NLP algorithms are employed to analyze the semantic and syntactic structure of students' answers, allowing for the identification of key concepts, logical coherence, and grammatical accuracy.

Complementing NLP, OCR technology plays a crucial role in AASES by facilitating the extraction of textual information from handwritten or printed answer scripts. OCR systems utilize image processing algorithms to recognize and convert text from scanned documents into machine-readable format. This capability is particularly valuable in educational settings where answer scripts may be handwritten, ensuring that AASES can evaluate responses across a variety of formats.

The integration of NLP and OCR technologies in AASES offers several benefits over traditional manual evaluation methods. Firstly, it significantly reduces the time and effort required for grading, enabling educators to focus more on providing personalized feedback and guidance to students.

Moreover, AASES can handle large volumes of answer scripts with consistency and impartiality, minimizing the impact of subjective biases inherent in human grading. Additionally, by providing instantaneous feedback, AASES promote active learning and encourage students to reflect on their responses, thereby fostering a deeper understanding of the subject matter.

Furthermore, AASES have the potential to adapt and evolve over time through continuous learning and refinement. By analyzing patterns in students' responses and feedback from educators, these systems can enhance their accuracy and effectiveness, ultimately leading to improved assessment outcomes. Moreover, AASES can be customized to accommodate different evaluation criteria, curriculum requirements, and language variations, making them versatile tools for educators across various disciplines and educational settings.

Despite the numerous advantages offered by AASES, challenges remain in their implementation and deployment. Ensuring the accuracy and reliability of NLP and OCR algorithms, especially in handling diverse languages, handwriting styles, and contextual nuances, is critical to the success of these systems. Additionally, addressing concerns related to data privacy, security, and ethical considerations is paramount to building trust and acceptance among stakeholders. Nonetheless, with ongoing advancements in NLP and OCR technologies, coupled with concerted efforts in research and development, AASES are poised to revolutionize the educational assessment landscape, offering scalable, efficient, and objective means of evaluating students' performance.

2. DETAILED SURVEY
[1] The research question under study focuses on designing an effective, accurate automatic grading system, with marginal percentage error, for the Generally Theory-based subjects, having no disparity with the grading system used by educators. The response to this research question is the bottleneck in the manipulation of answer scripts, which results to increased time consumption, lack of efficiency, and most importantly, biases in the score assignment. The approaches used in this paper incorporate Natural Language Processing (NLP), semantic analysis, and ontology with the aim of creating intelligent

grading system. In turning the answer scripts into machine-readable format, the OCR feature is adopted in this system for identifying textual content alone, but is also capable of dealing with other components such as tables and figures. The paper presents method of and best approach to grading using machine learning techniques as well as the application of support vector machines in grading. Unfortunately, the usually involved datasets in the respective decade are not discussed from the search results in the paper. However, it is probable that the researchers employed a set of answer scripts for training/development and a set of grading criteria/rubrics for grading the students' papers.

[2] acknowledges effectiveness of an automatic method of essay scoring in mitigating the issue of limited time in marking writing assignments and subjectivity of the grading process. The procedures adopted in the foregoing paper involve the use of Natural Language Processing (NLP), sentiment analysis, and machine learning specifically the Long Short-Term Memory (LSTM) models for essay grading where the essays are written in English. The phenomena are identified using NLP algorithms and it utilizes syntactic, semantic and sentiment features of the essays to predict the grades by employing LSTM models.

[3] The paper identifies the challenges and limitations of manual evaluation of subjective answers, such as bias, inconsistency, time consumption, and human resources. It aims to develop a system that can automate the evaluation process and reduce the need for human intervention. The paper presents a two-part system: a checker and an evaluator. The checker takes a question, a student's answer, an expected answer, and total marks as input, and assigns a score to the student's answer based on grammar, keywords, and similarity. The evaluator takes a sample of student's answers and finds the best combination of evaluation techniques and weights for each question. The system allows the user to choose from different methods for keyword extraction, summarization, and similarity check, or use the optimal combination suggested by the evaluator.

[4] The paper is organized by first presenting the background work which is divided into the research techniques which include the similarity measures and the machine learning techniques. The paper also overviews the pros and cons of the methods and offers some recommendations for an ideal grading system: You can see that the automation of the answers valuation scripts provides the grading system bias-free and coherent. This is why there is a need to establish a model that erases the precisions and achievements the grading performance because the outcome of the assessments concerns the student's future. Consequently, having reviewed the study, the authors established that there exists two primary strategies in answer grading; similarity measures and Machine Learning strategies. While similarity-based measures do not require a large training set, these methods are not effective in cases where it needs to mine for open-ended responses. On the other hand, ML techniques expanded the possible coverage of grading systems and they do well even with the semi-open-ended questions. This means an enormous labelled training set is needed to solve each question which may not be convenient at all.

[5] The paper uses various methodologies for each component of the system, such as OCR, NLP, machine learning, and similarity algorithms. For image text extraction, the paper uses py-tesseract, which is a Python-based OCR tool that converts images into text. For summarization, the paper uses a keyword-based technique that selects the most frequent words and avoids the less frequent words to generate a summary. For text preprocessing, the paper uses NLTK, which is a popular framework for natural language processing, and performs tokenization, stop-word removal, lemmatization, bigram creation, and word frequency count. For information retrieval, the paper uses a word2vec model to convert words into vectors and measure their semantic similarity. For mark scoring, the paper uses four similarity measures: cosine similarity, Jaccard similarity, bigram similarity, and synonym similarity, which compare the student's answer with the correct answer and calculate a score based on the angle, intersection, structure, and synonyms of the sentences.

[6] proposes a system that consists of the following steps: input image, preprocessing, feature extraction, text recognition, NLP techniques, data splitting, classification, mark evaluation, and performance metrics. The system uses the py-tesseract library for OCR, the mean and standard deviation for feature extraction, the artificial neural network (ANN) for classification, and the number of words and letters for mark evaluation. The paper uses various methodologies such as image processing, OCR, NLP, and deep learning to implement the proposed system. The paper also uses some tools such as tkinter, matplotlib, and numpy for data handling and visualization. It reviews some of the previous works related to OCR, NLP, and answer evaluation using machine learning. The paper cites some of the challenges and limitations of the existing methods and highlights the novelty and advantages of the proposed system.

[7] Preprocessing the answer scripts involves data processing including methods like tokenization, lemmatization, and word embedding which converts the answer scripts into numerical vector form. To do so, the paper employs deep learning techniques such as LSTM, Recurrent neural networks, and dropout and other methods to learn the semantic representation of the answer scripts and then assign score to them. In the study, the D-DAS is trained and evaluated through a supervised learning technique by providing answer scripts along with human-assessed scores as the manual dataset. The paper looks at the existing literature on AES, and other short answer grading systems, culminating in their strengths and weaknesses. It also walks through various forms of LSTM models, including simple LSTM, deep LSTM, and bidirectional LSTM, as well as their use in practical natural language processing and information retrieval applications.

[8] The paper also outlines the various earlier works done on the use of a computer to evaluate, text mining and measurement of text similarity. For the assessment of the student performance, there currently exists an evaluation paradigm which involves a powerful and effective Natural Language Processing (NLP) algorithm. This research was followed by the creation of the tool that incorporates the NLP analysis along with the Artificial Neural Network (ANN) to perform calculations. A filter set for matching an answer to the examination process is developed by the faculty in form of an answer sheet and a keyword dataset corresponding to the answer for the examination process. In this context, these datasets are contained in a data storage system. The results are then compared to the ANN algorithm to identify if they contain the correct answer from the student. Also, the student's answer is corrected for spelling and grammatical mistakes whenever there is unevenness using the NLP algorithm. The results generated from the text mining technique are calculated as soon as the techniques from NLP and ANN reach the end of their process.

[9] presents NLP techniques, such as tokenization, part-of-

speech tagging, stop word removal, stemming, and semantic similarity checking, to preprocess and analyze the student answers and compare them with the standard answers. Latent Semantic Analysis (LSA), which is an NLP technique based on a mathematical model that creates a vector representation of a document and measures the similarity between documents by calculating the distance between vectors. Bilingual Evaluation Understudy (BLEU), which is an algorithm that analyzes and measures the similarity between the student answer and the standard answer based on the n-gram co-occurrence matching procedure.

[10] presents a system for online paper evaluation using NLP for handwritten answer sheets and automatic mark sheet publishing. The system consists of the following modules: registration and login, upload, OCR, tokenization, similarity check and scoring. The system allows students to upload their scanned answer sheets and teachers to upload their answer keys. The system then converts the answer sheets into text using OCR, tokenizes the text and removes stop words, compares the text with the answer keys using WordNet and Corpus, and assigns marks based on the cosine similarity measure. The system also generates a mark sheet for each student and displays the results to the users. The paper also uses the cosine formula to calculate the similarity score between the answer sheet and the answer key, and to determine the marks obtained by the student.

[11] In this paper, the use of NLP and ML in creating a model to assess free-response answer scripts. This paper shall attempt at offering a solution to the general problem of the way in which answer scripts in formative and summative assessments to general tests and examinations are evaluated, especially during the COVID-19 pandemic and the lockdown. Accordingly, the paper presents a model responsible for the scoring of descriptive answers with the help of the similarity feature that can be calculated with the help of answer keywords extracted from the reference solution. The paper also examines several prior systems and research studies that deal with the issue of using text perception assessment for the assessment of the answer scripts by employing text extraction, similarity estimation, BLEU engineering modification, probabilistic semantic/text relatedness assessment, ontology, artificial neural network, Wordnet, Word2vec, WMD, cosine similarity, multinomial naïve Bayes, and term frequency-inverse document frequency. The paper validates if the system works according to the model on a local dataset by comparing the reference answers with student answers on the computerized tests and comparing the two sets of answers on a manual basis. The paper states that the proposed model was able to get average accuracy of 80% and developed a text file that gives the score for the answers. To support their arguments, the paper also presents a graphical representation of the validation process carried out manually and with the proposed system.

[12] The paper proposes a system called Automatic Answer Checker (AAC), which consists of a web-based interface for uploading question papers and answer sheets, and a machine learning module for analyzing and scoring the answers. The system uses natural language processing techniques such as word tokenization, stop- word and punctuation removal, and stemming to preprocess the text and extract keywords. The system then compares the keywords in the student's answer with the keywords in the model answer and calculates a similarity score. Based on the score, the system assigns marks to the student and displays them on the web interface.

[13] for the provision of a system for the automatic scoring of descriptive answers of machine learning. Feature extraction is another important process involved in the system where features are extracted through n-gram, cosine similarity, latent semantic analysis, and string similarity. It is also employed in the use of categorization models such as artificial neural networks, support vector machines, and linear regression to assign grades. The system also affords giving the specific scores reflecting the level of answers, recommendations and tips. This paper highlights the literature review focusing on the implicit automated question answering natura language, and the evolution of research in this field starting from the initial advancement in artificial intelligence till the present time. The paper categorizes the existing systems into three types: which are called corpus-based, information extraction, and mapping. Furthermore, the paper also provides an overview of the research limitations and future tasks in the domain which include content analysis, semantic analysis, and feedback system.

[14] addresses the challenge of evaluating students' performance through answer scripts. Traditional manual evaluation can be biased and is influenced by various factors like the mood swing of the evaluator and the inter-relation between the student and evaluator. The paper proposes an automatic answer script evaluation system based on Natural Language Processing (NLP). The system takes a student's written answer as input and automatically scores marks after the evaluation. The system considers all possible factors like spelling error, grammatical error, and various similarity measures for scoring marks. The system uses NLP for handling the English language used in the answers. For summary generation from the extracted text, keyword-based summarization techniques are used. Four similarity measures (Cosine, Jaccard, Bigram, and Synonym) are used as parameters for generating the final mark. The paper discusses the motivation behind the automated answer script evaluation, which includes less time consumption, less manpower involvement, prohibiting human evaluator's psychological changes, and easy record keeping and extraction.

[15] The paper presents a text analysis pipeline consisting of four stages: OCR, sentence boundary detection, tokenization, and part-of-speech tagging. The paper uses freely available opensource software packages for each stage, and applies them to a large dataset of scanned news articles with different levels of degradation. It then compares the results of the text analysis stages on the clean and noisy versions of the same documents using the proposed evaluation paradigm, which can identify and track individual OCR errors and their cascading effects. The paper also proposes a novel evaluation paradigm based on hierarchical dynamic programming to measure and analyze the impact of OCR errors on NLP stages.

3. ARCHITECTURE

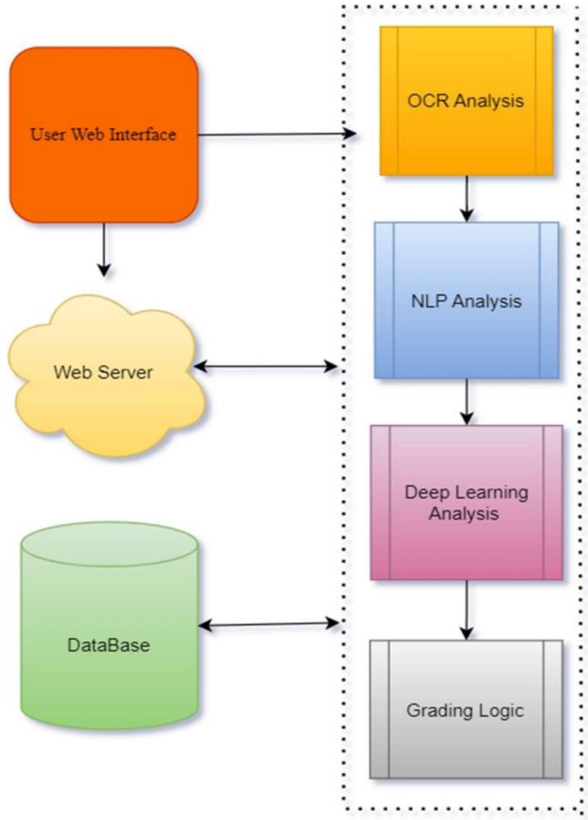

Fig 1. Proposed Architecture

The suggested architecture of the system offers a complete solution to automate the process of checking answer scripts with the use modern technologies for efficiency and precision as well as ensuring that user friendliness and data security is maintained. A web-based interface that is easy to navigate for both teachers and learners takes center stage in this architectural design. Educators may upload the scripts, view evaluated results and give feedback through this hub. It is designed in such a way that anyone can understand how it works easily thus allowing them to interact with different parts seamlessly.

Another important integration involves an Optical Character Recognition (OCR) system. This component makes it possible to extract textually based information from responses including those written by hand or containing non-textual features thereby setting ground for further examination. Next, written responses are analyzed by Natural Language Processing (NLP) algorithms which consider their semantic content and coherence. NLP analysis investigates language subtleties, measures comprehension depth and checks contextual appropriateness. The program also uses sophisticated linguistic processing methods to determine student response quality more accurately.

The architecture is supported by a safe database system so that the answer scripts, OCR results, NLP analyses, grades and feedback can be stored securely, in compliance with privacy regulations and ensuring confidentiality as well as integrity. This strong backend infrastructure serves as the spine of this system where it also protects sensitive data while enabling various functions to take place. In general terms then; proposed structure represents an all-round, advanced technological approach towards streamlining work-flows during assessment automation while at the same time improving on educational experiences among teachers as well learners.

4. METHODOLOGY

4.1 Data Collection and Preprocessing

1) Answer Script Collection: Collect a diverse set of handwritten or typed answer scripts from various educational institutions or examinations. Ensure that the dataset covers a range of subjects, difficulty levels, and writing styles.

2) Digitization: Scan the collected answer scripts to create digital images or documents that can be processed by the OCR system.

3) Ground Truth Preparation: Establish a ground truth dataset by manually grading a subset of the collected answer scripts. This ground truth will be used to train and validate the NLP algorithms.

4.2 OCR Processing

1) OCR Implementation: Implement an Optical Character Recognition (OCR) system to extract the textual content from the digitized answer scripts. Ensure that the OCR system can handle both textual and non-textual elements (e.g., diagrams, formulas) present in the answer scripts.

2) OCR Accuracy Evaluation: Assess the accuracy of the OCR system by comparing the extracted text with the ground truth data. Identify and address any issues or limitations in the OCR performance.

4.3 NLP Analysis

4) Feature Extraction: Develop NLP algorithms to extract relevant features from the OCR-processed text, such as semantic content, language complexity, coherence, and contextual relevance.

5) Scoring Model Development: Design a scoring model that can effectively evaluate the quality and correctness of the written responses based on the extracted features. Incorporate techniques like text similarity, sentiment analysis, and knowledge-based scoring.

6) Model Training and Validation: Train the scoring model using the ground truth dataset. Employ cross-validation techniques to ensure the model's generalization and robustness.

7) Model Optimization: Continuously refine and optimize the NLP algorithms and scoring model based on the performance on the validation dataset.

4.4 Data Storage and Management

1) Database Design: Design a secure database system to store the digitized answer scripts, OCR results, NLP analyses, grades, and feedback.

2) Data Integrity and Privacy: Ensure data integrity, confidentiality, and compliance with relevant privacy regulations throughout the data storage and management processes.

3) Database Integration: Integrate the database seamlessly with the other components of the proposed system, enabling efficient data storage, retrieval, and management.

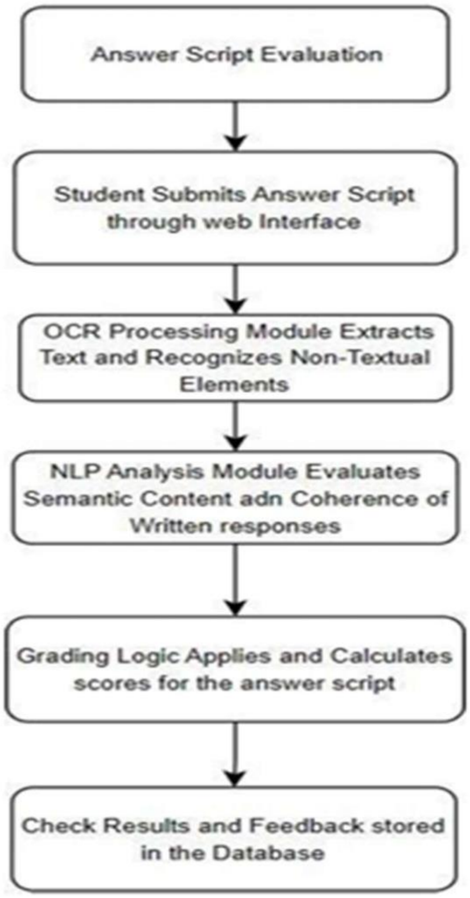

Fig 2. Workflow Diagram

5. RESULTS

To orchestrate this sophisticated system, a methodology was devised to modernize the assessment process of educational institutions.” A dynamic website with strong login authentication is built to upload and view answer scripts using the above steps. The site also has an intuitive interface that can be easily explored, the treasure trove of digitized scripts can be accessed by student ID, department, semester exam and subject.

Upon submission, the answer scripts are subjected to an OCR process that converts the handwriting or typed data into computer-readable text form. But this is not a simple mechanical conversion of text — it’s really the beginning of exactly what is often need: a place where key constructs from each response are built, recorded and available for analysis.

The extracted text is carefully recorded in a secured database for further evaluation and feedback from levels of human control later on. But the magic of the system is in its NLP capabilities, where algorithms have been trained to hone in on the language and actually read though those answers. These algorithms are very good at picking up on semantic nuances, peeling back the layers of complexity and checking responses for coherence and consistency along many dimensions. With this linguistic expertise, they construct a bespoke scoring framework that ensures the questions are assessed fairly and thoughtfully. Moreover, even the evaluation is generated, the response is looked from cosine similarity point of view to be exactly matched with some ground truth in training corpus. This analysis is the objective foundation for awarding marks in a manner that guarantees fairness and alleviates any teacher bias in grading. Meanwhile, for answers adorned with diagrams and visual representations, a cutting-edge deep learning model takes center stage. This model, finely attuned to the intricacies of visual data, delivers a nuanced assessment based on the similarity and fidelity of diagrams, enriching the grading process with a holistic perspective.

In essence, this amalgamation of OCR, NLP, and deep learning technologies heralds a new era in educational assessment, one characterized by precision, transparency, and adaptability. The website stands not only as a testament to technological innovation but also as a beacon of progress, ushering in a paradigm shift in the way academic achievement is perceived and evaluated. With data integrity and privacy enshrined at its core, this system embodies the ideals of trust and accountability, paving the way for a future where assessment transcends mere scrutiny, evolving into a catalyst for growth and excellence.

6. CONCLUSION

The development and implementation of an Automated Answer Script Evaluation System represent a pivotal advancement in the educational technology landscape, aiming to address the challenges associated with manual evaluation processes. The system outlined in this report integrates cutting-edge technologies such as Optical Character Recognition (OCR) and Natural Language Processing (NLP) to revolutionize the grading paradigm. The comprehensive set of functional requirements, usability enhancements, and non-functional considerations collectively shape a robust framework for an efficient, accurate, and user-friendly solution.

The system's key functionalities, including user authentication, answer script submission, OCR processing, NLP analysis, non-textual element recognition, grading interface, feedback mechanism, and data storage, collectively ensure a holistic approach to automated evaluation. By implementing role-based access control and real-time feedback mechanisms, the system not only streamlines the evaluation process but also contributes to improved educational outcomes and personalized learning paths.

The emphasis on non-functional requirements, including performance, scalability, usability, maintainability, and compatibility, underscores the commitment to delivering a solution that meets the highest standards of efficiency, reliability, and adaptability. The software requirements, centered around web hosting, NLP modules, and a secure database, along with specific hardware prerequisites, form the backbone of a technology stack designed to handle the complexities of large- scale assessment processes.

7. REFERENCES

[1] A. Rokade, B. Patil, S. Rajani, S. Revandkar, and R. Shedge, "Automated Grading System Using Natural Language Processing," in 2018 Second International Conference on Inventive Communication and Computational Technologies (ICICCT), Coimbatore, India, 2018, pp. 1123-1127, doi: 10.1109/ICICCT.2018.8473170.

[2] V.S. Sadanand, K.R. Guruvyas, P.P. Patil, J. Janardhan Acharya, and S. Gunakimath Suryakanth, "An automated essay evaluation system using natural language processing and sentiment analysis," International Journal of Electrical and Computer Engineering (IJECE), 2022.

[3] V. Kumari, P. Godbole, and Y. Sharma, "Automatic Subjective Answer Evaluation," 2023, doi: 10.5220/0011656000003411.

[4] A.K.R. Maya, J. Nazura, and B.L. Muralidhara, "Recent

Trends in Answer Script Evaluation – A Literature Survey," 2023, doi: 10.2991/ahis.k.210913.014.

[5] Prof. S.P. Raut, S.D. Chaudhari, V.B. Waghole, P.U. Jadhav, and A.B. Saste, "Automatic Evaluation of Descriptive Answers Using NLP and Machine Learning," 2022, doi: 10.48175/IJARSCT-3030.

[6] S. K. et al., "Automatic Answer Evaluation Using Deep Learning Algorithms," 2022, doi: 10.31838/ecb/2023.12.s3.039.

[7] G. Ng'Ochoi, P. Sijimol, and S. Mariam Varghese, "Grading descriptive answer scripts using deep

[8] learning," International Journal of Innovative Technology and Exploring Engineering, vol. 8, pp. 991-996, 2019.

[9] V. Lakshmi and V. Ramesh, "Evaluating Student's Descriptive Answers Using Natural Language Processing and Artificial Neural Networks," ISSN: 2320-2882, 2017.

[10] N. D. Kamraj, "Survey on Techniques used for Evaluation of Exam Answer Papers," 2020.

[11] M. Sebastian, R. Kunjumon, S. Shaji, and Prof. S. R., "DigiValuate: Answer Sheet Evaluation System using Natural Language Processing," 2021.

[12] S. Mangesh, P. Maheshwari, and A. Upadhyaya, "Subjective Answer Script Evaluation using Natural Language Processing," 2022.

[13] V. Tanwar, "Machine Learning based Automatic Answer Checker Imitating Human Way of Answer Checking," INTERNATIONAL JOURNAL OF ENGINEERING RESEARCH & TECHNOLOGY (IJERT), vol. 10, no. 12, December 2021.

[14] 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," in 2020 International Conference on Inventive Computation Technologies (ICICT), Feb. 2020, pp. 439–443, doi: 10.1109/ICICT48043.2020.9112429.

[15] M. M. Rahman and F. H. Siddiqui, "NLP-based Automatic Answer Script Evaluation," 2018.

[16] D. Lopresti, "Optical character recognition errors and their effects on natural language processing," IJDAR, vol. 12, pp. 141–151, 2009, DOI: 10.1007/s10032-009-0094-8.