Automatic Grading of Answer Sheets using Machine Learning Techniques

Kasarapu Ramani[1*] Guggilla Uma Maheswari[2], Kattamanchi Prem Krishna[3] Sagabala Venkata Meghashyam[4] , Komirisetty Venkata Pavan Kumar[5] , Yuvaraj Duraiswamy[6]

[1] Professor& Head, Department of CSE(DS), Mohan Babu University (Erstwhile Sree Vidyanikethan Engineering College), Tirupathi, India
[2, 3, 4, 5] UG Scholar, Department of Computer Science and Systems Engineering, Sree Vidyankethan Engineering College, Tirupati, India.
[6]Professor, Department of Computer Science, Chan University, Duhok, Iraq

*ramanidileep@yahoo.com umaguggila22@gmail.com premkrishna.kattamanchi@gmail.com shyamsagabala03@gmail.com,

sumanthpavan70714@gmail.com, d.yuvaraj@duhok.edu.krd

Abstract. Automating the grading process for question-answer sheets represents a significant challenge, particularly when dealing with traditional hard copy papers. This initiative aims to reduce the time and expenses associated with manual grading, a task that typically consumes 2-3 days for teachers to complete. Leveraging advanced Natural Language Processing (NLP) and Machine Learning (ML) methodologies, including XGBoost, Ridge Regression, and Naive Bayes, we have developed a system for automatic grading using prepossessed OCR datasets. This system learns from a historical dataset of student question-answers, with a focus on two primary objectives: scoring short-answer questions and providing constructive feedback to students. Additionally, we assess the performance and accuracy of the system using standard evaluation techniques such as Precision, Recall, and F-measure. Our experimental results demonstrate an impressive 89% accuracy in grading student answer sheets.

Keywords: NLP,Machine learning, Naïve Bayes, OCR, XGBoost,. Ridge, Regression and Performance Metrics.

1 Introduction

In the field of education, grading question-answer sheets is still a labor-and time-intensive process that requires a lot of physical labor from teachers. But now that cutting-edge advances in machine learning and natural language processing are accessible,

K. R. Madhavi et al. (eds.), *Proceedings of the International Conference on Computational Innovations and Emerging Trends (ICCIET 2024)*, Advances in Computer Science Research 112,
https://doi.org/10.2991/978-94-6463-471-6_27

there's a chance to totally change this procedure. This research aims to tackle this problem by automating the grading of hard copy question-answer sheets using advanced NLP and ML algorithms.

The most objective of this research work is to reduce the time and expenses related to manual grading, which takes two to three days for teachers to do on average. By employing methods such as XGBoost, Ridge Regression, and Naive Bayes, the system seeks to deliver precise evaluations of student responses, relieving educators of the laborious hand grading duty.

The implementation of NLP and ML algorithms is made possible by the use of optical character recognition technology, which transforms hard copy question-answer sheets into machine-readable format. The system attempts to understand patterns and subtleties present in responses by utilizing historical datasets of student question-answers, with a major focus on scoring short-answer questions.

Beyond just assigning grades, the system also seeks to give students insightful feedback that will expedite their learning and support their academic growth. Strict criteria like Precision, Recall, and F-measure are utilized to assess the system's performance, guaranteeing the precision and dependability of the automated grading procedure.

All things considered, this paper is a big step toward improving efficacy and efficiency in the field of education, demonstrating how cutting-edge technologies may simplify routine tasks and give teachers and students more control.

2 Literature Survey

A survey was conducted on various research papers to comprehend different techniques and highlight the significance of this work. Here, a few of them are being discussed.

Sanuvala & Fatima[1] The introduced Handwritten Answer Evaluation System as a solution to streamline the grading process of student exam papers. By utilizing Optical Character Recognition technology and machine learning/Natural Language Processing techniques, HAES aims to automate and standardize the assessment procedure. This method offers advantages such as efficiency in reducing manual effort and ensuring consistency in grading criteria. However, challenges such as potential inaccuracies in OCR and subjectivity in model training may affect the system's accuracy and fairness. Overall, HAES represents a significant step towards improving the efficiency and reliability of evaluating handwritten exam responses in education.

Harsh Jain et al.[2] introduced an innovative approach to evaluating theory-based exam papers, combining Natural Language Processing techniques with Optical Char-

acter Recognition technology.This approach involves converting handwritten responses into digital text and utilizing vector embeddings to compare them with teacher-provided solutions, enabling automated grading and ensuring consistency in evaluation standards. However, it acknowledges the potential for inaccuracies in OCR and the complexities associated with implementing NLP, which could impact grading precision and necessitate careful consideration. Despite these challenges, the system represents a significant leap forward in automating the assessment of handwritten exam answers, offering both efficiency gains and the need for further refinement to address accuracy concerns.

David Becerra-Alonso et al. [3] introduced EduZinc, a comprehensive tool empowering teachers in the creation and evaluation of course materials. This model facilitates two main functions: Firstly, it aids in crafting individual learning products such as activities, exams, and tests. Secondly, it automates the grading process for these products, along with generating student profiles, classes, and reports. Additionally, the system incorporates automatic notifications, alerting both struggling and excelling students accordingly. However, the model's accuracy performance was hindered by its lack of integration with classification technology for assigning final marks to students.

Liu et al. [4] present an innovative method for extracting and understanding global features from short text. Through a fusion of Convolutional Neural Networks (CNN) and Latent Dirichlet Allocation (LDA), the authors devised a technique capable of capturing both local and global features effectively. This approach holds substantial promise for various natural language processing tasks, especially in the realm of short text comprehension.

Neslihan et al. [5] conducted research focusing on the UK's GSEC exam, offering feedback to students. Initially, the authors employed standard data mining techniques to analyze student answers alongside model responses. They subsequently devised similarity measures based on common word occurrences, utilizing a clustering algorithm for this purpose.

The new system brings in improvements to streamline the grading of question answer sheets. By using NLP and ML methods such as XGBoost, Ridge Regression and Naive Bayes it effectively assesses student answers, cutting down on grading time substantially. Additionally, besides assigning scores to responses the system also offers feedback to students enriching their learning journey. After assessment using metrics, its effectiveness is confirmed, showcasing a significant progress, in grading efficiency and precision.

3 Proposed Methodology

This paper uses three different algorithms: XGBoost, Ridge Regression, and Naive Bayes to evaluate question-answer sheets autonomously. Each algorithm is unique and provides a fresh approach to solving problems associated with assessing students' responses.

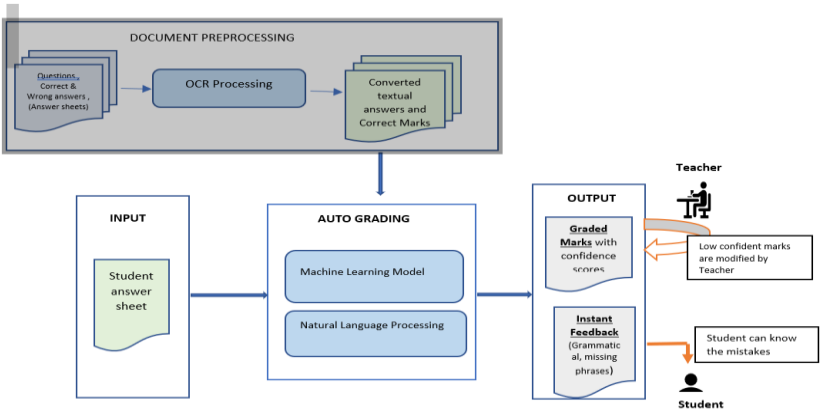

Fig1: Architecture of the Automatic grading of answer sheets

The Auto-grading system (AGS) depicted in Fig.1 is composed of multiple machine learning and natural language processing components

3.1 Data Collection and Preprocessing:

The first phase of the project involved the collection of question-answer sheets from educational institutions. These sheets were obtained in both digital and hard copy formats to ensure the diversity of data sources. Additionally, metadata such as subject, grade level, and type of questions were recorded for each dataset. We use the publicly available high school student question-answer pairs dataset from the internet for building a learning-based system.Student answer sheets available in internet.

Once collected, the question-answer sheets underwent extensive preprocessing to ensure uniformity and compatibility for further analysis. And the data is checked manually whether the data is accurate to the responses entered by the students.

Feature Extraction:
The next step involved extracting relevant features from the pre-processed text data. Features such as word frequency, sentence structure, and semantic meaning were identified and extracted using techniques from Natural Language Processing. Additionally, domain-specific features relevant to the subject matter were considered to enhance the accuracy of the grading process.

Model Training and Evaluation:

A variety of machine learning algorithms were considered for automating the grading process, including XGBoost, Ridge Regression, and Naive Bayes. These algorithms were evaluated based on their performance on a training set, considering metrics such as accuracy, precision, recall, and F1 Score. Following evaluation, the selected models were trained using the pre-processed dataset of student responses.

3.2 Derivation of Algorithms

In this paper, we undertake the task of automating the grading process for question-answer sheets by employing three distinct algorithms: Naive Bayes, XGBoost, and Ridge Regression. Each algorithm brings its unique strengths to the table and it is executed in the Microsoft VS Code Editor offering diverse approaches to tackle the complexities inherent in assessing the Dataset of 10,000 samples containing question, student answer and awarded marks are collected from the schools. Currently it is available for the project.

Naive Bayes

The straightforwardness and effectiveness of the traditional probabilistic classifier Naive Bayes are well-known. When it comes to text classification tasks, it frequently performs extremely well despite its "naive" assumption of feature independence. In our implementation, As mentioned in Fig 2 we utilize the Multinomial Naive Bayes variant, which is particularly suitable for discrete features such as word counts or TF-IDF vectors. By modeling the probability distribution of features given class labels, Naive Bayes provides a straightforward yet effective method for grading student responses.

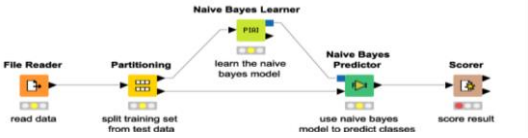

Fig 2: working of naïve bayes algorithm

XGBoost:

In Fig 3 as classified XGBoost, or eXtreme Gradient Boosting, has emerged as a state-of-the-art algorithm for supervised learning tasks. By employing an ensemble of decision trees and optimizing a customizable loss function, XGBoost achieves remarkable accuracy and scalability. In our context, XGBoost's ability to handle large datasets and complex relationships between features and scores makes it a compelling choice for automating the grading process. By iteratively refining predictions through the boosting technique, XGBoost offers a powerful tool for accurately assessing student responses.

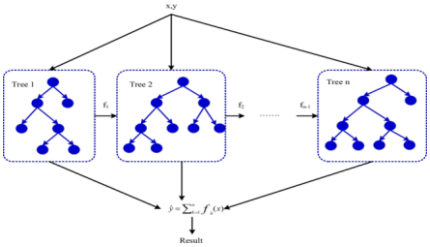

Fig 3: implementation of XGBooster classifier

Ridge Regression:

Ridge Regression is a linear regression technique that addresses the issue of multicollinearity and over-fitting by introducing a regularization term to the cost function. By penalizing large parameter values, Ridge Regression produces more stable and interpretable models, particularly in high-dimensional datasets. In our implementation, Ridge Regression serves as a complementary approach to Naive Bayes and XGBoost, providing a linear model that captures the underlying relationships between features and scores in a concise manner.

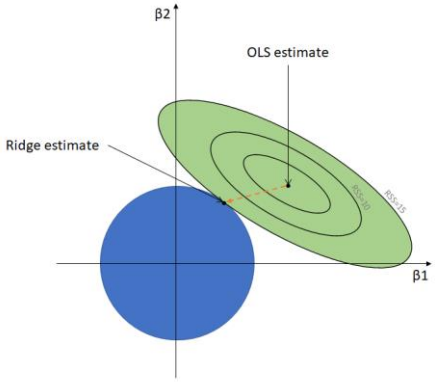

Fig 4.representation of ridge regression model

Implementation Steps:

Our implementation proceeds through several key steps:

Input: We start by reading the training data, comprising questions, reference answers, student responses, and corresponding scores. We then extract features from the textual components using appropriate techniques such as TF-IDF vectorization or simple concatenation.

Model Training: For Naive Bayes, we discretize scores into bins and vectorize features using TF-IDF. We then train a Multinomial Naive Bayes classifier on the transformed features. For XGBoost and Ridge Regression, we prepare feature matrices and corresponding score vectors and train the models on the feature-score pairs, optimizing hyperparameters as necessary.

Grading: Utilizing the trained models, we predict scores for student responses. For Naive Bayes, we transform features using the pre-trained TF-IDF vectorizer and predict

scores using the classifier. For XGBoost and Ridge Regression, we directly predict scores based on the learned models.

Evaluation: At last, we assess the performance of each show utilizing suitable measurements such as accuracy, precision, recall, and F1 score. By comparing the comes about, we point to distinguish the foremost compelling calculation for reviewing question-answer sheets.

4 Results

The research introduces a solution for automating the assessment of question answer sheets specifically focusing on copies to reduce the time and expenses linked with grading. By applying Natural Language Processing and Machine Learning methods like XGBoost, Ridge Regression and Naive Bayes the system uses an OCR dataset generated from hard copy papers. It is tailored to grade short answer questions and offer feedback to students by referencing a database of student responses. Performance evaluation metrics such, as Accuracy, Precision, Recall and F measure are used to gauge the effectiveness of the system. The experimental results were analyzed to identify trends, strengths, and limitations of the proposed system. Comparative investigation was performed to compare the execution of each calculation and distinguish the foremost viable approach for computerizing the evaluating prepare. The implications of the findings were discussed, highlighting the potential impact of automated grading systems on educational practices and student outcomes. The below are the mathematical equations of Accuracy, Precision, Recall, F-measure i.e., Eqn - 1 to Eqn - 4 respectively.

$$Accuracy = \frac{(True\ positives + True\ Negatives)}{Total\ no\ of\ Test\ samples} \quad \text{-------- (1)}$$

$$Precision = \frac{(True\ positives)}{True\ Positives + False\ positives} \quad \text{-------- (2)}$$

$$Recall = \frac{(True\ positives)}{True\ Positives + False\ Negatives} \quad \text{-------- (3)}$$

$$F - measure = 2X \frac{(Precision * Recall)}{(Precision + Recall)} \quad \text{-------- (4)}$$

These metrics indicate the robustness and effectiveness of the XGBoost algorithm in accurately assessing student responses. With high precision and recall rates, XGBoost ensures both the accuracy and comprehensiveness of the grading process, highlighting its potential as a reliable tool for automating educational evaluations.

Table 1: Performance Comparison Among model-algorithm Naïve Bayes, Ridge Regression, XG Boost

Algorithm	Precision	Recall	F-Score
Naïve Bayes	0.893	0.701	0.785
Ridge Regression	0.923	0.794	0.854
XGBoost	0.948	0.850	0.897

Based on experimental results, XGBoost demonstrates superior efficiency in the automatic grading of question-answer sheets, achieving impressive performance metrics:

Fig 5: Comparison of models Accuracy

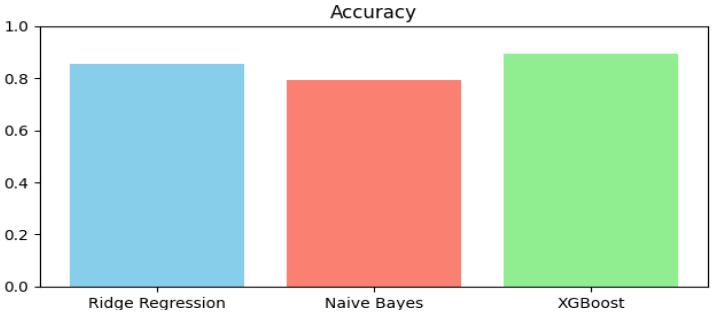

Fig 6: Comparison of models Recall

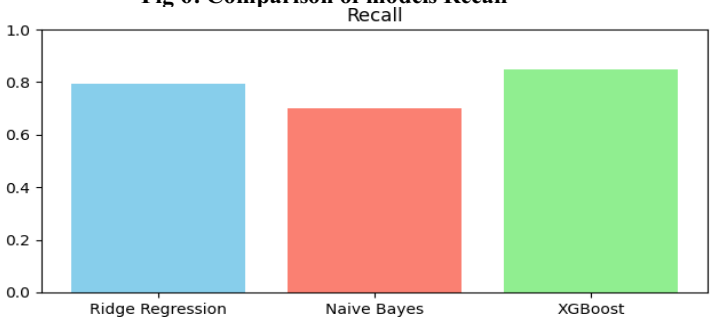

Fig 7: Comparison of models Precision

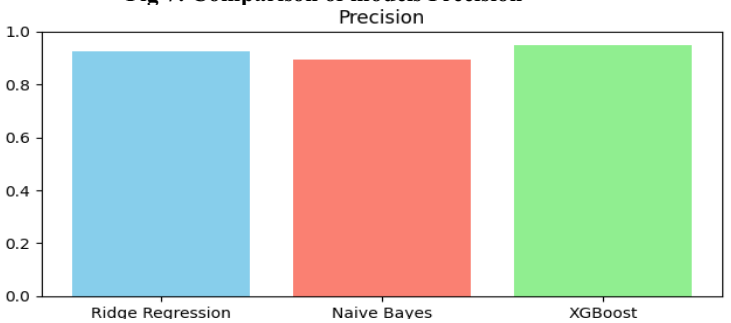

Fig 8: Comparison of models F1 Score

Comparison between the performances:
In our evaluation of Ridge Regression, Naive Bayes, and XGBoost for automating the grading process of question-answer sheets, we observe notable differences in their performance metrics. Assessment measurements such as Precision, Recall, and F-measure were utilized to degree the accuracy and reliability of the automated grading system.

5 Conclusion & Future work

In conclusion, this paper speaks to a critical walk in automating the grading process of question-answer sheets through the application of advanced Natural Language Processing and Machine Learning techniques. By employing algorithms such as Naive Bayes classification and Ridge Regression and XGBooster, we have developed a system capable of efficiently assessing student responses and providing valuable feedback, thereby alleviating the burden on educators and enhancing the overall efficiency of the educational evaluation process and experimental results reveal that XGBoost achieves the highest efficiency. Through rigorous evaluation metrics and comprehensive implementation, this project underscores the transformative potential of technology in streamlining traditional educational practices, paving the way for more effective learning outcomes and resource optimization in educational institutions.

In the future,Public datasets were available not too long ago, and research in the field is open to new techniques, However, it's still not widely used in automatic grading. Using technology like ICR to convert handwritten text into digital form can help. Then, applying deep learning techniques for grading can improve accuracy. Additionally, using LLMs can offer students personalized feedback, making learning better. This approach could greatly enhance automatic grading systems in education

References

1. Sanuvala, G., & Fatima, S. S. (2021). A Study of Automated Evaluation of Student's Examination Paper using Machine Learning Techniques. 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCIS).
2. Harsh Jain, Mohd SherAli Shaik, Ravi Shankar, Vinita Mishra:Student, Information Technology, VESIT, Mumbai, Maharashtra, India,Volume: 09 Issue: 05 | May 2022
3. Becerra-Alonso, D., Lopez-Cobo, I., Gómez-Rey, P., FernándezNavarro, F. and Barbera, E., 2020. EduZinc: a tool for the creation and assessment of student learning activities in complex open, online, and flexible learning environments. Distance Education, 41(1), pp.86-105.
4. J. Liu, H. Ma, X. Xie, and J. J. E. Cheng, "Short TextClassification for Faults Information of Secondary EquipmentBased on Convolutional Neural Networks," vol. 15, no. 7, p.2400, 2022.
5. Süzen, N., Gorban, A.N., Levesley, J. and Mirkes, E.M., 2020. Automatic short answer grading and feedback using text mining methods. Procedia Com- puter Science, 169,

pp.726-743.
6. Neethu George, Sijimol PJ, Surekha Mariam Varghese, "GradingDescriptive Answer Scripts Using Deep Learning", InternationalJournal of Innovative Technology and Exploring Engineering (IJITEE), Volume-8 Issue-5 March, 2019.
7. Alomran, M., & Chia, D. (2018). Automated Scoring System forMultiple Choice Test with Quick Feedback. International Journal of Information and Education Technology, 8(8).
8. Mohler, M., Bunescu, R., Mihalcea, R.: Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp. 752–762 , 2011
9. A. Vinaya Babu, "Detection of concept-drift for clustering time-changing categorical data: An optimal method for large datasets." In Data Engineering and Communication Technology: Proceedings of 3rd ICDECT-2K19, pp. 861-871. Springer Singapore, 2020.
10. Kuraparthi, Swaraja, Madhavi K. Reddy, C. N. Sujatha, Himabindu Valiveti, Chaitanya Duggineni, Meenakshi Kollati, and Padmavathi Kora. "Brain Tumor Classification of MRI Images Using Deep Convolutional Neural Network." Traitement du Signal 38, no. 4 (2021).
11. Magooda, A., Zahran, M.A., Rashwan, M., Raafat, H., Fayek, M.B.: Vector based techniques for short answer grading. In: International Florida Artificial Intelligence Research Society Conference Ahmed, pp. 238–243 (2016)
12. Passero, G., Haendchen Filho, A., Dazzi, R.: Avaliação do uso de métodos baseados em se wordnet para corre¸c˜ao de quest˜oes discursivas. In: Brazilian Symposium onComputers in Education (Simp´osio Brasileiro de Informática na Educa¸c˜ao-SBIE), vol. 27, p. 1136 (2016)
13. Ziai, R., Ott, N., Meurers, D.: Short answer assessment: establishing links between research strands. In: Proceedings of the Seventh Workshop on Building Educational Applications Using NLP, pp. 190–200, 2012.
14. Valenti, S., Neri, F., Cucchiarelli, A.: An overview of current research on automated essay grading. J. Inf. Technol. Educ. 2, 319–330, 2003