Volume 9, Issue 7, July – 2024 International Journal of Innovative Science and Research Technology
ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24JUL1710

Automated Grading Model with Adjusted Level of Lenience for Short Answer Questions using Natural Language Processing

S Zindove[1]; S Chaputsira[1]
School of Information Science and Technology, Harare, Zimbabwe [1]

Abstract:- Automated grading of short answer questions is a challenging task that involves understanding and evaluating free-text responses. This research presents an innovative model that combines the capabilities of the language model all-mpnet-base-v2 with a machine learning-based lenience adjustment mechanism to enhance the accuracy and fairness of automated grading systems. The proposed model utilizes all-mpnet-base-v2 for natural language understanding and feature extraction from student responses. To address the variability in acceptable answers and provide a fair grading system, a machine learning-based model is integrated to adjust the level of lenience dynamically. This dual approach ensures that the grading system can handle a wide range of responses while maintaining consistency and reliability. The experimental results demonstrate that the combination of all-mpnet-base-v2 with the lenience adjustment model significantly improves grading accuracy compared to traditional methods. This model represents a significant advancement in the field of educational technology, offering a robust solution for automated grading systems that can adapt to diverse educational contexts and requirements.

Keywords:- *All-Mpnet-Base-V2, Lenience, Convolutional Neural Networks, Pretrained Models.*

I. INTRODUCTION

The rapid advancement in Natural Language Processing (NLP) has revolutionized various fields, including education, where automated grading systems for short answer questions have become a significant area of research. Automated grading systems offer the potential to reduce the workload of educators, provide immediate feedback to students, and ensure consistency in grading. This paper explores an automated grading model that incorporates an adjusted level of lenience for short answer questions, leveraging the capabilities of the language model **all-mpnet-base-v2** and a machine learning-based lenience adjustment mechanism. The process of automated grading involves evaluating student responses and assigning grades based on the quality and accuracy of the answers. Traditional methods often rely on keyword matching and predefined rules, which can be rigid and fail to capture the nuances of human language. Recent advancements, such as the utilization of transformer-based models like BERT and its variants, have shown promise in understanding and evaluating natural language with higher accuracy (Kumar et al., 2019; Liu et al., 2019).

However, these models still face challenges, particularly in handling variations in student responses and ensuring fairness in grading. The need for a flexible and adaptive grading mechanism that can adjust its lenience based on the context and content of the answers is crucial. This paper proposes a novel approach that combines the semantic understanding capabilities of all-mpnet-base-v2 with a machine learning model designed to adjust grading lenience dynamically.

By employing all-mpnet-base-v2, a transformer-based model known for its superior performance in sentence embeddings and semantic similarity tasks, the proposed system aims to capture the intricacies of student responses more effectively. Furthermore, the machine learning model for lenience adjustment leverages historical grading data to fine-tune its parameters, ensuring that the grading process remains fair and consistent across different contexts and student demographics (Guerra et al., 2020; Lun et al., 2020).

The integration of these technologies represents a significant step forward in automated grading systems, providing a robust framework that can adapt to various educational settings and grading standards. This paper will delve into the architecture of the proposed model, the training and evaluation methodologies, and the results obtained from experimental studies, demonstrating its effectiveness and potential for widespread adoption in educational institutions.

In addition to improving grading accuracy, integrating a lenience adjustment mechanism addresses the critical issue of bias in automated grading systems. Bias in grading can stem from various factors, including the diverse linguistic backgrounds of students and differing interpretations of grading rubrics by individual educators. The proposed model's ability to dynamically adjust its lenience based on contextual understanding helps mitigate such biases, promoting a more equitable evaluation process. This adaptability is particularly important in diverse educational environments, where a one-size-fits-all approach to grading may not be feasible (Magooda et al., 2016; Pérez-Marín et al., 2009). Through rigorous testing and validation, this study aims to demonstrate the effectiveness of our model in delivering fair and reliable automated grading outcomes.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24JUL1710

II. BACKGROUND STUDY

Automated grading of short answer questions (SAQs) has garnered significant attention due to its potential to streamline the educational assessment process. Traditionally, grading has been a manual, time-consuming task that is prone to inconsistencies and biases. The integration of automated systems aims to address these challenges by providing quick, consistent, and objective grading.

The foundation of automated grading systems lies in comparing student responses to reference answers. Early approaches predominantly used rule-based and statistical methods, which relied heavily on lexical and syntactic features. For example, simple keyword matching and frequency-based techniques were employed to determine the similarity between student answers and the model answer. However, these methods were limited in their ability to understand context and semantics, leading to inaccuracies, especially in more complex and varied responses.

The advent of machine learning (ML) and natural language processing (NLP) technologies marked a significant advancement in automated grading. Machine learning algorithms, particularly Support Vector Machines (SVMs), were among the first to be applied to this problem. These algorithms classified responses based on features extracted from the text. Despite their Deep learning, and more specifically neural network-based approaches, brought a paradigm shift. Models like Long Short-Term Memory networks (LSTMs) and Convolutional Neural Networks (CNNs) allowed for the processing of sequential data and the capture of complex patterns within the text. Recent models like BERT (Bidirectional Encoder Representations from Transformers) and its variants have further enhanced the ability to understand context and semantics. These models leverage large-scale pre-training on diverse text corpora, enabling them to generate more accurate and contextually aware representations of student responses.

In addition to accuracy, the concept of leniency in grading has been a critical area of exploration. Leniency involves adjusting the grading criteria to account for acceptable variations in student responses, ensuring fairness and reducing the potential for penalizing minor deviations from the reference answers. Studies have shown that incorporating leniency mechanisms can significantly improve the reliability and educational value of automated grading systems.

III. RELATED WORK

Automatic grading of short answer questions (SAQs) has been an active area of research due to its potential to enhance educational assessment by providing quick, consistent, and objective grading. Several methodologies and techniques have been explored in this field, ranging from traditional statistical methods to advanced machine learning and natural language processing (NLP) techniques.

A. Traditional Approaches

Initially, many automatic grading systems relied on rule-based or statistical methods. These methods often used lexical and syntactic features to compare student responses to reference answers (van der Waa, J., Nieuwburg, E., Cremers, A. and Neerincx, M., 2021. Evaluating XAI: A comparison of rule-based and example-based explanations. Artificial intelligence, 291, p.103404). For example, techniques such as bag-of-words and n-grams were employed to capture the similarity between the responses and model answers. However, these methods faced challenges in handling the variability and nuances of natural language. (Moore, S., Nguyen, H.A., Chen, T. and Stamper, J., 2023, August. Assessing the quality of multiple-choice questions using gpt-4 and rule-based methods. In European Conference on Technology Enhanced Learning (pp. 229-245). Cham: Springer Nature Switzerland.)

B. Vector-based Techniques:

These techniques represent text as vectors in a multi-dimensional space and use distance measures to determine similarity. Magooda et al. applied vector-based methods improvements over rule-based systems, they still faced challenges in capturing the deeper semantic meaning of responses to grade short answers by comparing their semantic content with reference answers, providing a foundational approach to automated grading. In their paper Tan, H., Wang, C., Duan, Q., Lu, Y., Zhang, H., & Li, R. (2020). Automatic short answer grading by encoding student responses via a graph convolutional network. Interactive Learning Environments, 31(3),1636–1650. Tan H Wang developed a grading system using graph convolutional network. The results showed that their model outperformed rule based grading systems on SemEval dataset.

C. Semantic Similarity Measures

Mohler et al. (2011) developed a system that employs semantic similarity measures and dependency graph alignments to assess the relevance and correctness of student answers. Their approach integrates multiple graph alignment features with lexical semantic similarity measures, using machine learning techniques to improve accuracy. By aligning the dependency graphs of student answers with those of reference answers, the system can evaluate not just the words used, but the structural relationships within the sentences, providing a more comprehensive assessment of the answer's meaning.

D. Deep Learning Models

Liu et al. (2019) introduced multiway attention networks, which capture more nuanced features from student responses, outperforming traditional methods. These models utilize attention mechanisms to focus on relevant parts of the text, enhancing their ability to understand context and semantics. Researchers are applying Deep Learning Approaches for the past five years to address this problem owing to the increasing popularity of this area. The paper by Liu aims to summarize various existing deep learning approaches researchers followed to address this problem and to investigate whether Deep Learning based techniques are outperforming traditional approaches.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24JUL1710

E. Transformer-based Models

The use of transformer-based models like BERT has become prevalent due to their ability to handle complex language tasks. These models excel in understanding context and nuances in text, making them highly suitable for applications in natural language processing (NLP). One significant application of BERT is in educational technology, where it can assist in automated grading and feedback systems and reasoning provided by students, which are crucial for accurate assessment. By identifying and evaluating the justifications, BERT can provide more nuanced and precise grading compared to traditional keyword-based methods. Moreover, the application of BERT in this context goes beyond mere grading. It has the potential to offer detailed feedback to students, highlighting areas where their justifications were strong and pointing out gaps where they could improve. This can enhance the learning experience by providing personalized and constructive feedback, which is often challenging to deliver in large-scale educational settings.

The effectiveness of BERT in such tasks underscores its versatility and power. It can be trained and fine-tuned on specific datasets to cater to various educational needs, from automated essay scoring to evaluating open-ended questions. As educational institutions continue to adopt digital tools, the integration of advanced models like BERT can lead to more efficient, scalable, and fair assessment methods, ultimately contributing to improved educational outcomes.

➢ *Incorporating Leniency in Grading*

The integration of leniency in automated grading systems has been explored through various methods:

- *Machine Learning-Based Lenience Models*

These models adjust lenience dynamically based on historical grading data and contextual understanding. Kumar et al. (2019) developed AutoSAS, which includes mechanisms to adjust scoring based on question context and difficulty, providing a more balanced evaluation

- *Threshold-based Adjustments*

Reimers and Gurevych proposed setting thresholds for semantic similarity to adjust lenience. Their model graded responses more leniently if they closely matched the reference answer but contained minor deviations, improving the flexibility of the grading system.

- *Domain Adaptation*

Thakur et al. emphasized the importance of adapting grading models to specific subjects or types of questions. This approach fine-tunes lenience parameters to match educational objectives and ensures that grading remains fair across different contexts.

- *Heuristic Methods*

McDaniel et al. discussed heuristic methods that adjust scores based on response complexity and relevance. These methods aim to balance strictness with educational benefits, ensuring that students are not unfairly penalized for minor variations Ichikawa et al. utilized BERT for estimating justification cues in student answers, demonstrating its effectiveness in grading short answers by identifying key justifications within the text. This approach leverages BERT's contextual understanding to discern relevant information.

IV. METHODOLOGY

In our experiment to develop an automated grading model for short answer questions (SAQs) with adjusted leniency, we followed a structured approach encompassing data collection, pre-processing, model training, and leniency adjustment.

A. Data Collection and Pre-processing

➢ *Data Collection*

We collected a comprehensive dataset of SAQs and corresponding student responses. To evaluate our method for short answer grading, we created a data set of questions from Software Project Management assignments with answers. The assignments were administered as part of a Software Project Management course at the Harare Institute of Technology. The students submitted answers to 100 questions spread across 5 assignments. Table 1 shows two question-answer pairs with three sample student answers each. Thirty-one students were enrolled in the class and submitted answers to these assignments. The data set we work with consists of a total of 31 X 100 student answers (3100). The answers were independently graded by three lecturers, using a provided marking guide. We also gathered pre-marked answer scripts (P) from a subset of students to establish a baseline for leniency adjustments.

➢ *Data Pre-processing*

The collected data was cleaned by removing irrelevant information and standardizing the format of the answers. This involved lowercasing text, removing punctuation, and handling common text anomalies. We tokenized the text using the tokenizer provided by the Hugging Face library for the all-mpnet-base-v2 model. The dataset was then split into training, validation, and test sets, ensuring that the pre-marked answer scripts were included in the training set.

B. Similarity Calculation using all-mpnet-base-v2

➢ *Model Setup*

We loaded the pre-trained all-mpnet-base-v2 model from the Hugging Face library, which is known for its capability to capture semantic similarities and handle complex language tasks. The model was fine-tuned on our training dataset to adapt it specifically for the task of grading SAQs.

➢ *Encoding Responses*

Reference answers **(M)** and student responses **(R)** were encoded into dense vector representations using the all-mpnet-base-v2 model. Each reference answer **M= {m1, m2...}** and each student response **R={r1, r2,…}** were represented as vectors.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24JUL1710

➢ *Similarity Calculation*

Cosine similarity was calculated between each pair of encoded vectors from **M** and **R** to determine the degree of similarity. The cosine similarity score SSS for each pair was computed as follows:

$$S(M,R) = \frac{\sum_{i=i}^{n} m_i r_i}{\sqrt{\sum_{i=1}^{n} m_i^2}} \sqrt{\sum_{i=1}^{n} r_i^2}$$

These similarity scores provided a quantitative measure of how closely a student response R matched the reference answer M, considering both lexical and semantic content.

C. Leniency Calculation and Application

➢ *Pre-Marked Answer Script Analysis*

Pre-marked answer scripts **P** were analyzed to understand the grading patterns of human evaluators. This analysis included calculating average similarity scores $\boldsymbol{S_p}$ for these scripts and comparing them with their assigned grades. Statistical analysis was performed to identify patterns and establish a baseline for leniency adjustments, including metrics such as mean, standard deviation, and inter-rater reliability.

➢ *Leniency Model*

A leniency adjustment model was developed using machine learning techniques. This model incorporated features such as similarity scores $\boldsymbol{S_p}$, answer length, presence of key terms, grammatical correctness, and contextual relevance. We used a regression-based approach to train the leniency model on the pre-marked scripts ***P,*** enabling it to predict leniency-adjusted scores ***L*** for new responses. The model was optimized using cross-validation techniques to prevent overfitting and ensure generalizability.

➢ *Applying Leniency to Student Scripts*

For each student response ***R***, the initial similarity score ***S*** was calculated using the all-mpnet-base-v2 model. This score was then adjusted using the trained leniency model to produce the final grade. The adjusted score LLL was calculated as: ***L = f (S, features)*** where ***f*** represents the leniency model trained on the pre-marked scripts ***P***.

D. Final Grading

The final leniency-adjusted scores L were used to grade the remaining student's scripts. These scores were validated against human-graded examples in the test set to ensure accuracy and fairness.

Table 1: A Sample Question with Short Answers Provided by Students and the Grades Assigned by the Three Human Judges

	Sample questions, correct answers, and student answers	**Lecturer 1 Grade**	**Lecturer 2 Grade**	**Lecture 3 Grade**
Question 1 **Correct answer** **Marks**	What is the primary objective of project management in software development? Delivering the project on time, Staying within budget And Meeting the project's scope and quality requirements **3**			
Student answer 1 Student answer 2 Student answer 3	The primary objective of project management in software development is to ensure that the project is completed on time, within budget, and meets the specified requirements. To complete the project on schedule and within budget. The main goal is to finish the project successfully.	3 2 0	3 2 1	3 2 1
Question 2 **Correct answer** **Marks**	Describe the role of a project manager in a software project. Planning and defining project scope, Resource allocation, Risk management, Stakeholder communication and Monitoring and controlling project progress **5**			
Student answer 1 Student answer 2 Student answer 3	They manage the project and ensure it's on time. A project manager plans and defines the project scope, allocates resources, manages risks, communicates with stakeholders, and monitors and controls project progress. The project manager is responsible for planning, managing resources, and keeping the project on track.	1 5 3	2 5 4	1 5 5

V. EXPERIMENTAL RESULTS

A. Evaluation Metrics

The performance of the automated grading system was evaluated using multiple metrics: Mean Absolute Error (MAE) to measure the average magnitude of errors, Pearson correlation coefficient to assess the linear correlation between automated and human grades, and Quadratic Weighted Kappa (QWK) to measure agreement between the two sets of grades.

B. Cross-Validation

We employed k-fold cross-validation (with k=5) to ensure the robustness and generalizability of the model. This involved partitioning the data into k subsets, training the model on k−1 subsets, and validating it on the remaining subset. This process was repeated k times, with each subset serving as the validation set once.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24JUL1710

C. User Interface

A user-friendly interface was developed for educators. This interface allows for the input of questions and student answers, viewing of graded results, and understanding of the grading rationale. The interface provides detailed feedback on each graded answer, highlighting areas of strength and improvement based on the model's analysis.

VI. RESULTS

A. Evaluation Metrics

The model's performance was quantified using three key metrics: Mean Absolute Error (MAE), Pearson correlation coefficient, and Quadratic Weighted Kappa (QWK). These metrics provided a comprehensive evaluation of the model's grading accuracy and consistency.

Table 2: Evaluation Results

Metric	Value
Mean Absolute Error (MAE)	0.15
Pearson Correlation	0.87
Quadratic Weighted Kappa (QWK)	0.83

The evaluation results for specific questions and student responses are summarized in the following tables, providing insight into the model's grading performance across different assignments and questions. Table 3 shows the Grading Consistency across Lecturers and Automated System. The Automated grading system shows how consistent it is with the human examiners in its grading. Table 2 shows that over ally the automated grading system managed to have a correlation of 0.87 with the manually graded scripts. Correlation was calculated between the average mark scored by human judges against the mark the automated grading system scored.

Table 4 illustrates the effect of applying leniency adjustments to the scores of student responses based on pre-marked answer scripts. Here is a detailed explanation of the columns in the table:

- **Student ID**: This column identifies individual students using anonymized IDs (S1, S2, etc.).
- **Original Score**: This column shows the initial score assigned to each student's response by the automated grading model before any leniency adjustments were applied. These scores are based on the similarity calculations between the student responses and the reference answers using the all-mpnet-base-v2 model.

Table 3: Grading Consistency across Lecturers and Automated System

Question	Lecturer 1 Grade	Lecturer 2 Grade	Lecturer 3 Grade	Automated Grade
Q1	8	7.5	8	7.5
Q2	7	6.5	7	6.5
Q3	9	8.5	8.5	8.5
Q4	6	6.5	7	7

Table 4: Leniency Adjustments based on Pre-Marked Scripts

Student ID	Original Score	Leniency-Adjusted Score	Difference
S1	7.0	7.5	+0.5
S2	6.5	7.0	+0.5
S3	8.0	8.2	+0.2
S4	7.5	7.8	+0.3

- **Leniency-Adjusted Score**: This column presents the scores after the leniency model has been applied. The leniency model was trained on pre-marked answer scripts to understand how human graders may show leniency towards certain answers. The adjusted score reflects these
- **Difference**: This column represents the numerical difference between the original score and the leniency-adjusted score. A positive difference indicates that the leniency model increased the student's score, suggesting that the original grading may have been too stringent according to the patterns learned from the pre-marked scripts.
- **Student S1:** The original score of 7.0 was adjusted to 7.5, indicating that the leniency model identified areas where human graders might have awarded higher points, resulting in a 0.5-point increase.
- **Student S2**: Similarly, the original score of 6.5 was increased to 7.0, reflecting a 0.5-point adjustment. This suggests a consistent pattern where slight leniency is applied to responses that initially received lower scores.
- **Student S3**: The original score of 8.0 saw a smaller adjustment to 8.2, with a difference of 0.2 points. This indicates that higher-scoring answers required less adjustment, possibly because they were already close to the human graders' standards.

The leniency model effectively adjusts scores to better match human grading tendencies, ensuring fairness and consistency in the automated grading process. By incorporating leniency adjustments, the model acknowledges and corrects for potential biases in the initial automated grading, leading to final scores that are more representative of human judgment. This approach enhances the reliability

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24JUL1710

and acceptance of the automated grading system in educational settings.

Table 5: Evaluation Metrics Before and After Applying Lenience Adjustments

Metric	Value	Value after applying Lenience
Mean Absolute Error (MAE)	0.15	0.10
Pearson Correlation	0.87	0.94
Quadratic Weighted Kappa (QWK)	0.83	0.93

Table 5 compares the performance of the automated grading model using three metrics: Mean Absolute Error (MAE), Pearson Correlation, and Quadratic Weighted Kappa (QWK). The values before and after applying lenience adjustments are presented, highlighting the improvements in the model's accuracy, consistency, and agreement with human grading standards after lenience is applied.

- **Mean Absolute Error (MAE):** A lower MAE indicates better predictive accuracy. Before Lenience: The original MAE of 0.15 suggests that, on average, the model's predictions are 0.15 points away from the actual scores.
- **After Lenience:** The MAE drops to 0.10, showing an improvement in the model's accuracy. The lenience adjustment has reduced the average prediction error, making the model's scores closer to the actual scores.

➢ *A higher Pearson Correlation Indicates Better Alignment Between Predicted and Actual Scores.*

- **Before Lenience:** The original correlation of 0.87 indicates a strong positive linear relationship between the predicted and actual scores.
- **After Lenience:** The correlation increases to 0.94, demonstrating an even stronger relationship. The lenience adjustment has enhanced the consistency between the model's predictions and the actual scores.

VII. CONTRIBUTIONS

This research makes several significant contributions to the field of automated grading and natural language processing (NLP). These contributions enhance the understanding and application of automated systems for educational assessment, particularly in the grading of short answer questions (SAQs).

A. Development of a Leniency Adjustment Mechanism:

A novel leniency adjustment mechanism is introduced, which uses pre-marked student scripts to calibrate the grading system. This mechanism ensures that the automated grades are not only accurate but also fair, reflecting the leniency typically applied by human graders.

This approach helps address the inherent variability in human grading, making the automated system more robust and aligned with human judgment.

B. Integration of Transformer Models for Automated Grading:

This study demonstrates the application of the all-mpnet-base-v2 model, a state-of-the-art transformer-based model, in grading short answer questions. The use of such advanced models showcases their capability to understand and evaluate complex language tasks, contributing to more accurate and reliable automated grading systems. Ichikawa et al. utilized BERT for estimating justification cues in student answers, highlighting the effectiveness of transformer models in educational contexts (Ichikawa et al., 2020).

VIII. CONCLUSION

This research presents a significant advancement in the field of automated grading by integrating state-of-the-art natural language processing (NLP) techniques and developing a novel leniency adjustment mechanism. The key contributions of this study include the application of the all-mpnet-base-v2 transformer model for grading short answer questions (SAQs) and the introduction of a leniency adjustment framework based on pre-marked student scripts.

The methodology demonstrated that leveraging transformer models enhances the accuracy and reliability of automated grading systems. By employing semantic similarity measures and deep learning approaches, our model can effectively evaluate the underlying meaning of student responses, moving beyond surface-level text comparison. This approach addresses the limitations of traditional grading methods, which often struggle with the variability and nuances of natural language.

The implementation of a leniency adjustment mechanism further improves the fairness and consistency of automated grading. By calibrating the model with pre-marked student scripts, the system can emulate the leniency typically applied by human graders. This adjustment ensures that the automated grades are not only accurate but also equitable, reflecting a human-like grading standard.

Our evaluation on a dataset of 3100 student responses from a Software Project Management course at the Harare Institute of Technology demonstrated substantial improvements in key metrics. The Mean Absolute Error (MAE) decreased from 0.15 to 0.10, Pearson Correlation increased from 0.87 to 0.94, and Quadratic Weighted Kappa (QWK) improved from 0.83 to 0.93 after applying leniency adjustments. These results indicate that the leniency-adjusted model provides more accurate, consistent, and fair evaluations of student answers.

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24JUL1710

In conclusion, this research contributes to the broader field of educational technology by offering a robust framework for automated grading systems. The integration of advanced NLP models and fairness mechanisms ensures that these systems can be trusted by educators and students alike. Future work could explore further enhancements in leniency adjustments and the application of similar models to other types of educational assessments. The ongoing development and refinement of such technologies hold promise for significantly improving the efficiency and effectiveness of educational assessment processes.

REFERENCES

[1]. Baker, R., & Smith, L. (2019). Evaluating the fairness of automated grading systems: Bias, transparency, and explainability. Journal of Educational Technology Research, 35(2), 123-140.

[2]. Burrows, S., Gurevych, I., & Stein, B. (2015). The eras and trends of automatic short answer grading. International Journal of Artificial Intelligence in Education,

[3]. Chen, L., & He, Z. (2013). A machine learning based approach for automatic short answer grading. Proceedings of the 2013 International Conference on Artificial Intelligence, 534-539.

[4]. Gao, Y., & Zhu, J. (2021). Enhancing short answer grading with transformers and knowledge distillation. IEEE Transactions on Learning Technologies

[5]. Hijikata, Y., & Matsushita, K. (2017). A survey of natural language processing techniques for automatic short answer grading. Journal of Information Processing

[6]. Ichikawa, H., Fujii, H., & Tokunaga, T. (2020). Estimating justification cues in student answers using BERT for automatic grading. Proceedings of the Annual Meeting of the Association for Computational Linguistics

[7]. Liu, H., Luo, C., & Zhu, Y. (2019). Multiway attention networks for automatic grading of student essays.IEEE Transactions on Learning Technologies

[8]. Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2016). Intelligence Unleashed: An argument for AI in Education. Pearson Education

[9]. Magooda, A., Farag, M., & Hussein, M. (2019). Automatic short answer grading using semantic similarity measures. Computers & Education

[10]. Mohler, M., Bunescu, R., & Mihalcea, R. (2011). Learning to grade short answer questions using semantic similarity measures and dependency graph alignments. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies

[11]. Moore, S., Nguyen, H. A., Chen, T., & Stamper, J. (2023). Assessing the quality of multiple-choice questions using GPT-4 and rule-based methods. In European Conference on Technology Enhanced Learning

[12]. Nielsen, R. D., Ward, W., & Martin, J. H. (2008). Annotating students' understanding of science concepts. Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC

[13]. Phandi, P., Chai, K. M. A., & Ng, H. T. (2015). Flexible domain adaptation for automated essay scoring using correlated linear regression. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing

[14]. Pulman, S. G., & Sukkarieh, J. Z. (2005). Automatic short answer marking. Proceedings of the Second Workshop on Building Educational Applications Using NLP

[15]. Ramanathan, V., & Di Eugenio, B. (2014). Lightly supervised learning of procedural dialogue systems. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing

[16]. Riordan, B., & Klein, D. (2014). Unsupervised system for short answer grading using clustering. Proceedings of the 9th Workshop on Innovative Use of NLP for Building Educational Applications

[17]. Roy, D., & Roy, K. (2021). Short answer grading using machine learning: A survey.

[18]. Tan, H., Wang, C., Duan, Q., Lu, Y., Zhang, H., & Li, R. (2020). Automatic short answer grading by encoding student responses via a graph convolutional network. Interactive Learning Environments

[19]. van der Waa, J., Nieuwburg, E., Cremers, A., & Neerincx, M. (2021). Evaluating XAI: A comparison of rule-based and example-based explanations. Artificial Intelligence

[20]. Weston, J., & Hermann, K. M. (2015). Artificial Intelligence: Deep learning for answering

[21]. Zhang, Z., & VanLehn, K. (2016). Using learning technologies to support computer-based grading of student work. Journal of Educational Computing Research

[22]. Zhou, G., & Yang, M. (2017). Automatic short answer grading via multi-layered semantic matching. EEE Transactions on Knowledge and Data Engineering

[23]. Ramachandran, G., & Chakrabarti, A. (2020). Hybrid models for automatic short answer grading using NLP and deep learning. Journal of Educational Technology Development and Exchange (JETDE)

[24]. Saha, A., & Dey, L. (2012). Automatic grading of short descriptive answers in medical domain. Proceedings of the 13th International Conference on Intelligent Text Processing and Computational Linguistics

[25]. Silvestri, G., & Ferilli, S. (2013). Automatic grading of short student answers by semi-supervised short text clustering. Journal of Computing and Information Technology

[26]. Raman, M., & Yadav, N. (2022). Using machine learning for automated assessment of short-answer questions

ISSN No:-2456-2165 https://doi.org/10.38124/ijisrt/IJISRT24JUL1710

[27]. Farag, M., & Younis, M. (2018). Neural network-based methods for short answer grading: A survey. Information Processing & Management
[28]. Mutlu, E., & Aleven, V. (2012). Enhancing automated essay scoring with discourse structure and sentence specificity features. Journal of Educational Computing Research
[29]. Flor, M., & Futagi, Y. (2012). Automatic detection of preposition and determiner errors in ESL writing. Journal of Educational Computing Research
[30]. Yaneva, V., & Temnikova, I. (2017). Evaluating the readability of automatic short answer grading: A comparative study. Journal of Computing and Information Technology
[31]. Dascalu, M., & Trausan-Matu, S. (2014). Automatic feedback for improving student writing skills using linguistic features. Journal of Educational Technology & Society
[32]. Gao, Y., & Zhu, J. (2021). Enhancing short answer grading with transformers and knowledge distillation. IEEE Transactions on Learning Technologies,
[33]. Hijikata, Y., & Matsushita, K. (2017). A survey of natural language processing techniques for automatic short answer grading. Journal of Information Processing,
[34]. Ichikawa, H., Fujii, H., & Tokunaga, T. (2020). Estimating justification cues in student answers using BERT for automatic grading. Proceedings of the Annual Meeting of the Association for Computational Linguistics, 48-55.
[35]. Liu, H., Luo, C., & Zhu, Y. (2019). Multiway attention networks for automatic grading of student essays. IEEE Transactions on Learning Technologies,
[36]. Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2016). Intelligence Unleashed: An argument for AI in Education. Pearson Education.
[37]. Magooda, A., Farag, M., & Hussein, M. (2019). Automatic short answer grading using semantic similarity measures. Computers & Education, 129, 234-245.