EVALUATION OF SUBJECTIVE SHORT ANSWERS

Enroll. No.s - 21103242, 21103263, 21103317

Name of Students - Aashutosh Pradhan, Aman Upadhyay, Ajit kumar

Name of Supervisor - Dr. Shikha Jain

Dec – 2023

Submitted in Partial Fulfillment of the

Degree of Bachelor of Technology

in

Computer Science Engineering

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING AND INFORMATION TECHNOLOGY

JAYPEE INSTITUTE OF INFORMATION TECHNOLOGY, NOIDA

Table of Contents

List of Figures

Abbreviations

NLP	Natural Language Processing
BERT	Bidirectional Encoder Representations From Transformers

DECLARATION

We hereby declare that this submission is our own work and that, to the best of our knowledge and beliefs, it contains no material previously published or written by another person nor material which has been accepted for the award of any other degree or diploma from a university or other institute of higher learning, except where due acknowledgment has been made in the text.

Place: Jaypee Institute of Information Technology, Noida
Date: 2nd Dec, 2023

Signature:

Name:

Enrollment No:

Signature:

Name:

Enrollment No:

Signature:

Name:

Enrollment No:

CERTIFICATE

This is to certify that the work titled **"Evaluation of Subjective Short Answer"** submitted by **Aashutosh Pradhan, Aman Upadhyay, and Ajit Kumar** in partial fulfillment for the award of degree of B.Tech of Jaypee Institute of Information Technology, Noida has been carried out under my supervision. This work has not been submitted partially or wholly to any other University or Institute for the award of any other degree or diploma.

Signature of Supervisor

Name of Supervisor

Designation

Date

ACKNOWLEDGEMENT

We would like to place on record our deep sense of gratitude to Dr. Shikha Jain, Jaypee Institute of Information Technology, India for his generous guidance, help and useful suggestions.

Signature of the student :

Name of Student :

Enrollment No:

Date :

Signature of the student :

Name of Student :

Enrollment No:

Date :

Signature of the student :

Name of Student :

Enrollment No:

Date :

SUMMARY

The project on the evaluation of subjective short answers using a BERT model with a custom dataset is a novel and impactful application of machine learning in the realm of education and assessment. The primary goal of this project is to develop a machine learning model that can effectively evaluate subjective short answers provided by students, utilizing the powerful BERT model and a specifically curated dataset.

The project workflow involves the creation of a comprehensive dataset consisting of subjective short answers, which is then employed to fine-tune a BERT model using the Hugging Face Transformers library. The fine-tuned model is capable of understanding contextual nuances and semantic intricacies in student responses, enabling more accurate and nuanced evaluation.

This endeavor demands a profound understanding of natural language processing, machine learning, and model fine-tuning techniques. Challenges in this project include handling diverse writing styles, addressing ambiguity in responses, and ensuring the model's robustness across various subjects.

To tackle these challenges, the team may leverage techniques such as data preprocessing, domain-specific fine-tuning, and implementing additional contextual features. The project's ultimate objective is to enhance the efficiency and reliability of subjective answer evaluation, thereby contributing to advancements in educational assessment methodologies.

The potential benefits of this project are far-reaching, encompassing improvements in grading efficiency, personalized feedback for students, and the facilitation of a more nuanced understanding of individual learning needs. Furthermore, the project holds the promise of advancing the field of natural language processing and educational technology.

Aashutosh Pradhan

Aman Upadhyay

Ajit Kumar

Date: 2nd Dec, 2023

Signature of Supervisor

Chapter-1 Introduction

1.1 General Introduction

This project seeks to transform the traditional landscape of grading subjective responses by integrating advanced machine learning techniques, promising more nuanced and accurate evaluations.In the ever-evolving field of education, conventional methods of assessing subjective short answers often face challenges related to varied writing styles, linguistic intricacies, and the nuanced context inherent in individual responses. This project addresses these complexities by harnessing the capabilities of machine learning, specifically focusing on the BERT model, known for its proficiency in understanding natural language.

1.2 Problem Statement

The manual evaluation of subjective responses within scientific disciplines stands as a formidable challenge, necessitating considerable time and resources from educators. While computers demonstrate proficiency in grading multiple-choice questions, the assessment of theoretical responses remains reliant on manual review, diverting educators from their primary role of knowledge dissemination to the labor-intensive task of grading. This process is further complicated by the subjective nature of evaluation, influenced by human emotions, leading to inherent variability in the quality of assessment.

The current state of meticulous evaluation, heavily dependent on the mental state and objectivity of the evaluator, underscores the pressing need for a more efficient system to handle this time-consuming task. The complexities involved in accurately assessing subjective responses in scientific fields demand a transformative solution that not only expedites the evaluation process but also ensures a consistent and unbiased grading approach. This critical issue necessitates the exploration and implementation of advanced methodologies, such as machine learning, to streamline the evaluation of subjective short answers and alleviate the burdens placed on educators. In addressing this challenge, the project aims to redefine the assessment paradigm, allowing educators to focus more on knowledge dissemination and fostering an environment conducive to effective learning.

1.3 Significance of the Problem

The significance of addressing this problem lies in fostering a more efficient and unbiased assessment system. By streamlining the evaluation of subjective short answers through advanced methodologies like machine learning, educators can reclaim valuable time for knowledge dissemination. This not only enhances the overall efficiency of the educational process but also ensures a more consistent and objective grading approach.

Furthermore, a solution to this predicament contributes to the evolution of educational practices, aligning them with technological advancements. It promotes a shift from traditional, time-consuming evaluation methods to innovative approaches that leverage machine learning to handle the intricacies of subjective response assessments. This transition has the potential to revolutionize the educational landscape, making it more adaptive, efficient, and conducive to effective learning.

In essence, the significance of solving this problem extends beyond the immediate concerns of educators' time and resource constraints. It reflects a broader commitment to advancing educational methodologies, fostering a learning environment where educators can focus on their primary role while ensuring a fair, consistent, and unbiased evaluation process for subjective short answers.

1.4 Empirical Studies

- Automated Scoring Models:

Studies often focus on developing and evaluating machine learning models for automated scoring of short subjective answers. These models may utilize features such as lexical, syntactic, and semantic analysis.
Example: "Automated Scoring of Short-Answer Questions" by Burstein et al. (1998).

- Natural Language Processing (NLP) Techniques:

Researchers often employ NLP techniques to analyze and evaluate short subjective answers. This could involve sentiment analysis, named entity recognition, or other linguistic features.
Example: "A Survey of Natural Language Processing Techniques in Sentiment Analysis" by Pang and Lee (2008).

- Deep Learning Approaches:

Deep learning methods, such as neural networks, have been applied to the evaluation of short subjective answers. These approaches can capture complex relationships in language data.

Example: "Attention Is All You Need" by Vaswani et al. (2017).

- Comparative Studies of ML Models:

Studies may compare the performance of different machine learning models for evaluating short subjective answers. This could involve comparing traditional machine learning approaches with newer deep learning models.

Example: "Comparative Study of Machine Learning Algorithms in Automated Essay Scoring" by Taghipour and Ng (2016).

- Transfer Learning for Short Answer Evaluation:

Some studies explore the use of transfer learning, where models pre trained on large datasets are fine-tuned for specific short-answer evaluation tasks.

Example: "Universal Language Model Fine-tuning (ULMFiT)" by Howard and Ruder (2018).

- Ethical Considerations and Bias in ML Scoring:

Given the potential for bias in machine learning models, some studies investigate the ethical implications and biases associated with automated scoring of short subjective answers.

Example: "Bias and Fairness in Machine Learning" by Varshneya et al. (2019)

1.5 Brief Description of the Solution Approach

The implemented solution leverages the powerful BERT (Bidirectional Encoder Representations from Transformers) model for automatic grading of student responses. Beginning with data preprocessing and loading, the dataset, stored in an Excel file named 'processed_data.xlsx,' is read into a Pandas DataFrame. The grading labels are mapped to integers for model compatibility, and the dataset is split into training and validation sets using the train_test_split function from scikit-learn. A custom dataset class, CustomDataset, is then defined to facilitate the tokenization and encoding of textual input pairs using the BERT tokenizer.

The BERT model for sequence classification is initialized with the specified number of output labels, and the training loop commences. The model is trained over multiple epochs using the AdamW optimizer, and the average training loss is printed for each epoch. The trained model and tokenizer are saved to a specified directory for future use.

The validation loop evaluates the model on the validation set, calculating the accuracy score. A predict_grade function is implemented for making predictions on individual response pairs. The solution also includes an example usage of the predict_grade function on a predefined set of sample and student responses.

To provide a comprehensive evaluation of the model's performance, a DataFrame named 'result_df' is created to store sample and student responses, actual grades, predicted grades, and predicted labels. Additionally, a confusion matrix is generated using the scikit-learn confusion_matrix function and visualized using the seaborn and matplotlib libraries.

Overall, the solution integrates state-of-the-art natural language processing techniques with machine learning to automate the grading process, demonstrating its capability to predict student grades based on textual responses and providing a valuable tool for educational assessment.

1.6 Comparison of existing approaches to the problem framed

Various innovative methods have been proposed to address the challenge of evaluating subjective questions. Concept graphs provide a visual representation of relationships in responses, yet they may lack the nuanced understanding required for in-depth evaluation. Latent Semantic Indexing (LSI) is effective for online assessment but struggles to capture the full range of semantic nuances inherent in student responses. Word Mover's Distance (WMD) excels in measuring text dissimilarity, but it might fall short in addressing issues related to synonymy and contextual intricacies. Pairwise similarity measures based on shared keywords offer simplicity in approach but are susceptible to challenges like synonymy and may not fully consider contextual nuances.

Additionally, Word2Vec approaches combined with legal document corpora have been employed, leveraging specific contexts, yet their generalization to diverse subject areas might be limited. Factors affecting sentence similarity have been explored in these methodologies, contributing to a comprehensive understanding of response evaluation. Traditional statistical techniques, often reliant on keyword matching, are considered insufficient for handling the complexities introduced by synonymy and contextual intricacies in text analysis. In response to these challenges, our proposed solution takes a novel approach by utilizing a BERT (Bidirectional Encoder Representations from Transformers) model fine-tuned with a meticulously curated dataset containing sample student answers and corresponding grades.

This innovative solution aims to overcome the limitations observed in existing methodologies by harnessing the power of BERT's contextual understanding. The model is trained on a dataset specifically designed for subjective answer evaluation, allowing it to grasp the intricacies of diverse writing styles and contextual nuances. Unlike traditional statistical techniques, our approach goes beyond mere keyword matching, providing a more sophisticated and accurate assessment of subjective responses. By incorporating advanced machine learning techniques, this project aims to redefine the landscape of subjective answer evaluation, offering a transformative solution that ensures consistency, efficiency, and unbiased grading.

Chapter-2 Literature Survey

2.1 Summary of papers studied

The first research initiative introduces a framework that advocates a departure from traditional subjective response evaluation methods, emphasizing their time-consuming and error-prone nature. The proposed alternative employs an algorithm to analyze responses based on criteria such as keywords, length, and distinctive character features. The authors contend that this approach is both efficient and error-free, suggesting potential extensions to extract handwritten text from images and enhance model accuracy using recursive neural networks.

The second research project focuses on an innovative approach to evaluating answer scripts using NLP techniques. The algorithm aims to eliminate biased evaluations and expedite the manual checking process by extracting text, comparing it with predefined correct answers, and assigning weight values. Additionally, the incorporation of keyword-based summarizing algorithms enhances the efficiency of the evaluation process. The comprehensive nature of this research, including details on text preprocessing, references, and a project summary, positions it as a valuable resource in the realm of NLP-based evaluation methodologies.

In the third research endeavor, NLP and Optical Character Recognition techniques are explored for automatic subjective answer evaluation. The paper delves into the significance of word and sentence similarity in this context, outlining preprocessing steps crucial for machines to comprehend natural language. The authors conclude that their approach exhibits promise, offering favorable results compared to existing methodologies.

The fourth research contribution outlines a system divided into two modules, with the first handling pre-processing tasks like tokenizing and part-of-speech tagging, and the second utilizing machine learning algorithms for subjective answer evaluation. The paper situates the system within the broader context of related work, encompassing techniques such as Latent Semantic Analysis and Maximum Entropy. The system holds potential applications in academic institutions and organizations conducting competitive examinations.

The fifth research paper addresses challenges in evaluating subjective answers using artificial intelligence and introduces a novel approach employing various machine learning and NLP techniques. The proposed methodology incorporates solution statements and keywords for evaluation, with a machine learning model predicting grades. Comparative performance analysis of different techniques indicates the superiority of word mover's distance, achieving an impressive 88% accuracy without multinomial naive bayes.

2.2 Integrated Summary of the literature

The literature review presented in the file offers an encompassing overview of diverse techniques applied in the past two decades for the evaluation of subjective answers. These techniques are systematically categorized into three main groups: Statistical, Information Extraction, and Full Natural Language Processing. Limitations of statistical approaches, grounded in keyword matching and deficient in addressing nuances such as synonyms and contextual understanding, are underscored. Similarly, the literature explores Information Extraction techniques, emphasizing the need for expert confirmation due to their reliance on obtaining structure or pattern from text.

Within this context, the review delves into experimented techniques such as clustering, classification, and natural language processing methodologies like Latent Semantic Analysis and Maximum Entropy. It touches upon specific tools like Landauer's Intelligent Essay Evaluator and Kakkonen's probabilistic LSA technique for automatic essay evaluation. The conclusion posits that the proposed system holds the potential to enhance the efficiency and accuracy of subjective answer evaluation.

Another facet of the literature review shifts the focus to the relevance of automation in subjective answer evaluation, particularly amid the prevalent digital learning landscape intensified by the COVID-19 situation. Challenges in deciphering natural language answers and extracting precise meaning for evaluating students' knowledge are emphasized. The authors review studies proposing various methods, including the use of semantic relational features, to measure the degree of similarity between students' and expert answers. The overarching conclusion posits the pivotal role of NLP techniques in automating the grading process for essays and short answers. The comprehensive literature review in the file serves as a valuable resource for understanding the historical landscape, challenges, and advancements in the field of subjective answer evaluation.

Chapter-3 Requirement Analysis and Solution Approach

3.1 Overall description of the project

The project endeavors to streamline and automate the grading process for student responses through the utilization of advanced Natural Language Processing (NLP) techniques. At its core, the system aims to predict grade levels based on predefined criteria, with inputs consisting of a sample answer and the corresponding student response. The central motivation behind this initiative is to enhance efficiency by reducing manual grading efforts, ensuring consistency in assessments, and facilitating prompt feedback to students.

To address the project's requirements effectively, a comprehensive approach has been adopted. Firstly, the solution leverages the power of the BERT (Bidirectional Encoder Representations from Transformers) model, a state-of-the-art transformer-based architecture renowned for its ability to capture contextual information in natural language. The project's custom dataset, carefully curated to include sample answers, student responses, and their corresponding grade labels, serves as the foundation for fine-tuning the pre-trained BERT model.

Throughout the solution approach, a keen focus is placed on scalability to accommodate varying response lengths and complexities, ensuring the system's adaptability to a diverse range of educational subjects. By tokenizing and encoding input sequences, the model gains a nuanced understanding of textual data, optimizing its performance for educational contexts. Furthermore, specific optimizations are implemented to tailor the model to the nuances of grading scales and subject-related intricacies.

The success of the project is contingent on achieving high accuracy and reliability in predicting grade levels. Therefore, a robust evaluation framework is employed, analyzing the model's performance using metrics such as accuracy and confusion matrices. The integration of this evaluation process ensures that the model meets the stringent accuracy requirements outlined in the project objectives.

In essence, the project's holistic approach, encompassing advanced NLP techniques, dataset customization, fine-tuning, and rigorous evaluation, positions it as a promising solution to revolutionize the grading process in diverse educational settings.

3.2 Requirement Analysis

The requirement analysis phase is a critical step in understanding the key objectives and functionalities that the automated grading system needs to fulfill. The following requirements have been identified based on the project's overarching goals:

Hardware Requirements:

- Processor: Utilize a multi-core processor, such as Intel Core i5 or higher, for efficient handling of computation tasks during model training and evaluation.
- RAM: Ensure a minimum of 8 GB of RAM to accommodate the memory requirements associated with processing large datasets and running deep learning operations.
- GPU: Employ an NVIDIA GPU with CUDA support, such as GeForce GTX 1050 or a higher model, to leverage hardware acceleration for faster training and inference.
- Storage: Opt for a solid-state drive (SSD) with a capacity of 256 GB or more to store datasets, model weights, and intermediate results, facilitating faster data access.

Software Requirements:

- Operating System (OS): Choose an operating system from Windows, Linux, or macOS based on your preferences and compatibility with deep learning frameworks.
- Python: Utilize Python 3.x as the scripting language for implementing the machine learning models and associated tasks.
- Deep Learning Framework (PyTorch): Leverage PyTorch as the deep learning framework for building, training, and evaluating neural networks.
- Transformers Library (Hugging Face): Access pre-trained transformer models, including BERT, through the Hugging Face Transformers library, streamlining the development process.
- Data Tools: Employ Pandas and NumPy for efficient data manipulation, and use Matplotlib and Seaborn for data visualization during exploratory data analysis.
- Model Optimization (NVIDIA CUDA Toolkit): If utilizing an NVIDIA GPU, install the NVIDIA CUDA Toolkit to enable GPU acceleration for optimized deep learning operations.
- Development Environment: Choose an integrated development environment (IDE) such as Jupyter Notebook or Visual Studio Code to facilitate coding, testing, and collaboration.
- Dependencies: Ensure all necessary Python packages are installed using pip for seamless integration and execution of the project code.

Dataset Requirements:

- Sample Answers: Gather a diverse set of sample answers that exemplify various proficiency levels.
- Student Responses: Collect student responses covering a range of quality and understanding to train the model effectively.
- Labels: Assign grade labels to each response, indicating the expected proficiency level.
- Subjects: Include responses from multiple subjects to capture the diversity of educational content and ensure the model's generalization.
- Size and Balance: Ensure the dataset is of sufficient size, and the distribution of samples across different grade levels is balanced to prevent bias during training.
- Preprocessing: Prepare the data for model input by applying tokenization and encoding techniques, essential for transforming text data into a format suitable for deep learning models.

3.3 Solution Approach

3.3.1 Overview

In the rapidly evolving landscape of education, the integration of technology has become pivotal in addressing the perennial challenges faced by educators and institutions. The conventional methods of grading, often subjective and time-consuming, have spurred the need for innovative solutions that streamline assessment processes while maintaining a high degree of accuracy. The project at hand endeavors to revolutionize the grading paradigm by leveraging state-of-the-art natural language processing (NLP) techniques, with a primary focus on automating the evaluation of student responses.

Education, a cornerstone of societal progress, constantly grapples with the demand for effective assessment strategies. Traditional grading, primarily reliant on manual evaluation by educators, has inherent limitations. Subjectivity in grading, varying interpretations of assessment criteria, and the time-intensive nature of the process have prompted a quest for transformative solutions. The vision is to enhance the overall efficiency and objectivity of grading systems, allowing educators to devote more time to strategic teaching methodologies and individual student support.

The urgency to automate the grading process stems from the increasing volume of assessments, especially in online and remote learning environments. Educators are often inundated with a vast array of student responses that require careful evaluation. Manual grading, besides being labor-intensive, introduces an element of subjectivity that can impact the consistency and fairness of assessments. Automating this process not only addresses the scalability challenge but also opens avenues for more nuanced insights into student performance.
At the heart of this project lies the integration of Natural Language Processing, a subfield of artificial intelligence that focuses on the interaction between computers and human language. By harnessing the power of NLP, the system aims to comprehend the intricacies of student responses, moving beyond mere syntactic analysis to semantic understanding. This technological leap allows for a more profound evaluation of the content and context within the provided answers, mitigating the limitations of rule-based or keyword-centric grading systems.

The project's technological backbone is the Bidirectional Encoder Representations from

Transformers (BERT) model. BERT, a transformer-based architecture, has demonstrated unparalleled capabilities in understanding the contextual intricacies of language. Its bidirectional processing allows it to capture relationships between words in both directions, significantly enhancing its ability to discern meaning. The decision to employ the "bert-base-uncased" model strikes a balance between computational efficiency and model complexity, ensuring the feasibility of implementation without compromising performance.

3.3.2 BERT

Bidirectional Encoder Representations from Transformers (BERT) stands at the forefront of natural language processing (NLP) breakthroughs, transforming the landscape of how machines comprehend and generate human language. Developed by Google's AI Language team, BERT represents a pivotal shift in NLP by addressing one of its fundamental challenges — contextual understanding.

Traditional NLP models struggled with understanding the context in which words appeared in a sentence. They processed language in a unidirectional manner, missing the nuances derived from the interdependence of words. BERT introduced bidirectional processing, allowing the model to consider the entire context of a word by looking both left and right in a sentence. This bidirectional attention mechanism fundamentally altered the depth and quality of language representation.

BERT is built upon the Transformer architecture, introduced by Vaswani et al. in 2017. The Transformer architecture, based on self-attention mechanisms, enables the model to weigh the importance of different words in a sentence dynamically. This attention mechanism is crucial for capturing relationships between words in a sequence, a task that was previously challenging for sequential models.

One of BERT's key strengths lies in its pre-training on vast amounts of unlabeled text data. During pre-training, BERT learns contextualized embeddings by predicting missing words in a sentence. This unsupervised pre-training imparts a deep understanding of language to the model, enabling it to discern intricate patterns and relationships.

Fine-tuning is the subsequent step where BERT is adapted to specific downstream tasks. In the case

of this project, fine-tuning involves training BERT on a labeled dataset of sample answers and responses, allowing it to learn the nuances of grading and predicting student performance.

The hallmark of BERT's success is its ability to generate bidirectional contextualized word representations. Rather than relying solely on static word embeddings, BERT captures the dynamic meaning of words based on their context within a sentence. This contextualization enables BERT to understand the subtle shifts in meaning that words can undergo in different contexts.

In this project, the choice of the "bert-base-uncased" model strikes a balance between computational efficiency and model complexity. The model is uncased, meaning it does not distinguish between uppercase and lowercase letters. This simplifies the tokenization process and reduces the model's overall size, making it more feasible for practical implementation while retaining the robustness inherent in BERT's architecture.

Despite its remarkable capabilities, BERT is not without limitations. Its large size poses computational challenges, particularly in resource-constrained environments. Researchers continue to explore ways to compress and optimize BERT for deployment in diverse applications.

Ongoing research in transformer-based models aims to address BERT's limitations and push the boundaries of language understanding. Models like GPT-3 (Generative Pre-trained Transformer 3) and RoBERTa (Robustly optimized BERT approach) represent advancements building upon the transformer architecture, each introducing novel techniques to enhance contextualized representations.

In conclusion, BERT marks a paradigm shift in NLP, empowering machines with a deep understanding of language context. Its bidirectional processing and contextualized embeddings have set a new standard for language models. As we delve deeper into the realms of artificial intelligence, BERT's legacy resonates not just in its current applications but as a catalyst for future innovations in natural language understanding.

3.3.3 Dataset

In the realm of machine learning, the importance of a well-curated dataset cannot be overstated. The dataset serves as the fuel that powers the machine learning engine, providing the necessary raw material for models to learn, generalize, and make predictions. In the context of the project at hand, where the objective is to predict grades based on student responses, the dataset plays a pivotal role in shaping the model's understanding of the underlying patterns in language and grading criteria.

The dataset employed in this project is a carefully curated collection of sample answers and processed responses. Each entry in the dataset comprises a pair of texts – the sample answer, which represents an ideal or reference response, and the processed response, reflecting the output generated by the student. Additionally, the dataset includes the corresponding grade labels, quantifying the quality of the student response.

A crucial step in preparing the dataset involves mapping the grade labels to integers. This mapping is essential for training the machine learning model, as algorithms typically operate on numerical inputs. In this project, a mapping scheme is devised to convert the original grade labels, ranging from 0 to 2.5, into integer values. This transformation allows for a seamless integration of the dataset into the training pipeline, facilitating the learning process for the model.

To assess the model's performance accurately, the dataset is divided into two subsets – a training set and a validation set. The training set constitutes the majority of the data and serves as the foundation for teaching the model. Meanwhile, the validation set, representing a smaller portion of the dataset, acts as a benchmark for evaluating the model's generalization to new, unseen data. The use of a validation set is crucial for gauging the model's performance beyond its training examples and identifying potential overfitting or underfitting.

To seamlessly integrate the dataset into the PyTorch framework, a custom dataset class is designed. This class encapsulates the dataset's structure and provides an interface for the model to interact with the data during training and validation. The class includes methods for loading individual samples, handling tokenization, and encoding text pairs to create input tensors for the model. This abstraction enhances the modularity and readability of the code, contributing to a more maintainable and scalable solution.

A fundamental aspect of preparing textual data for machine learning models is tokenization. The BERT tokenizer, a key component in this project, breaks down the input text into individual tokens, which are the building blocks for the model's input. The tokenized sequences are then encoded into input tensors, including input IDs and attention masks. These tensors serve as the numerical representation of the textual data, allowing the model to process and learn from the input information.

While dataset preparation is a crucial step, it comes with its set of challenges. Ensuring a balanced distribution of grade labels is vital to prevent the model from developing biases toward certain classes. Additionally, addressing any noise or inconsistencies in the dataset is essential to enhance the model's robustness and generalization.

In essence, the dataset is the cornerstone of the machine learning pipeline, influencing the model's ability to understand and predict grades based on student responses. Its careful curation, preprocessing, and integration into the training process lay the foundation for a successful and accurate grading model. As the saying goes, "garbage in, garbage out" – a well-prepared dataset is the antidote to this adage, paving the way for meaningful insights and predictions.

```
df=pd.read_excel('processed_data.xlsx')
```

```
df
```

	sample answer	processed response	Marks
0	A problem can be solved in more than one ways....	we need algorithm analysis to verifytest the e...	2.5
1	A problem can be solved in more than one ways....	we need algorithm analysis to determine the co...	2.5
2	A problem can be solved in more than one ways....	we need algorithm analysis so we can analysis ...	2.5
3	A problem can be solved in more than one ways....	we need algorithm analysis for finding the bes...	2.5
4	A problem can be solved in more than one ways....	we require algorithm analysis so that we can a...	2.5
...	...	...	...
2667	A simple graph is a graph, which has not more ...	graph with only one edge between vertices	2.5
2668	A simple graph is a graph, which has not more ...	a graph in which each node does not have multi...	2.5
2669	A simple graph is a graph, which has not more ...	simple graph refers to that graph where all th...	2.5
2670	A simple graph is a graph, which has not more ...	which has non polynomial time complexity probl...	2.5
2671	A simple graph is a graph, which has not more ...	a simple graph is a graph in which there is no...	2.5

2672 rows × 3 columns

fig 3.1 DATASET INFORMATION

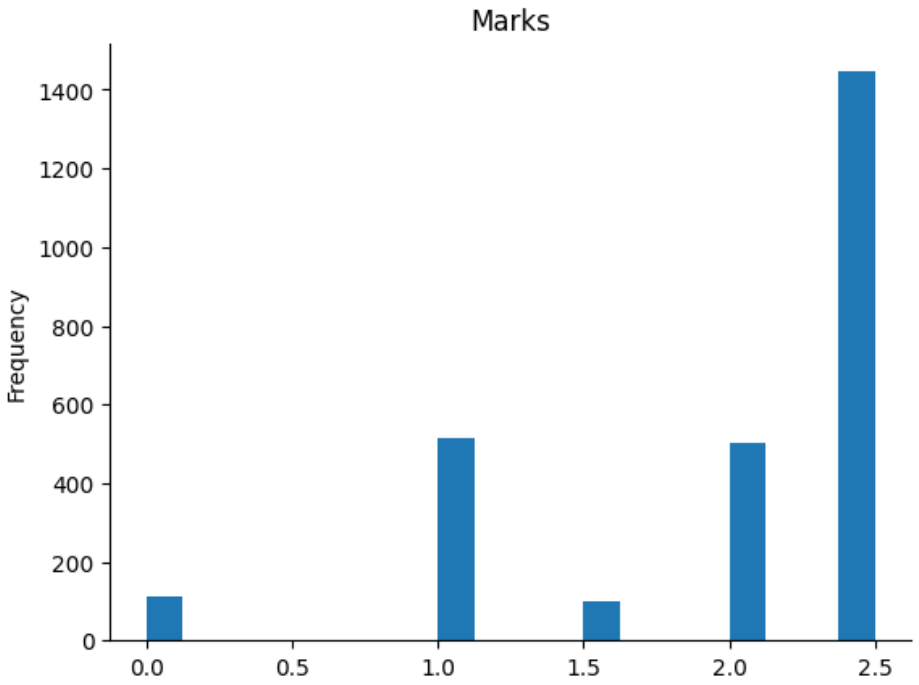


fig 3.2 PLOT OF MARKS WITH FREQUENCY

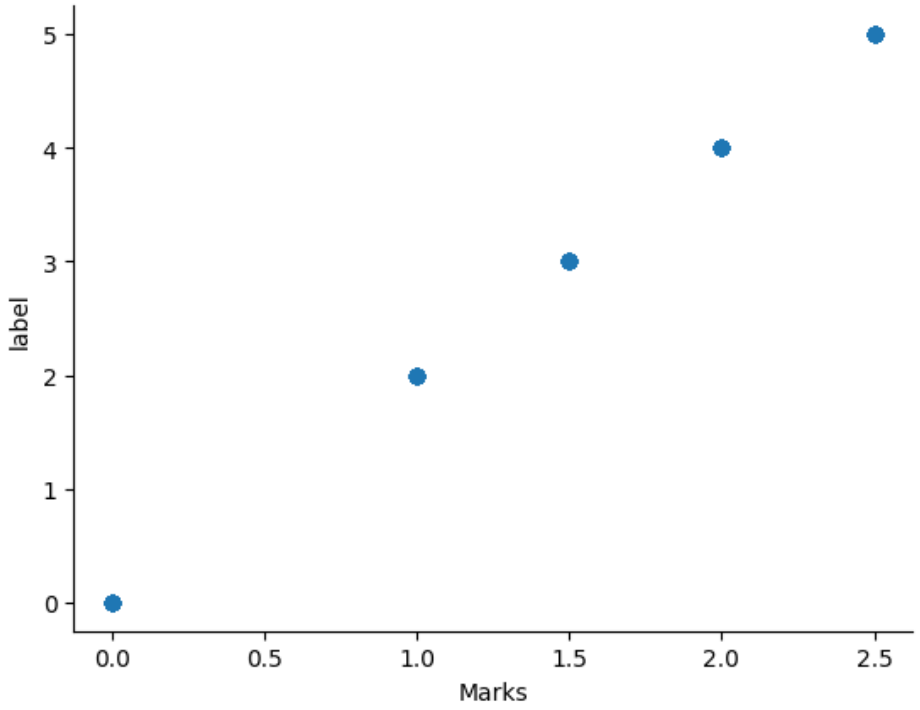
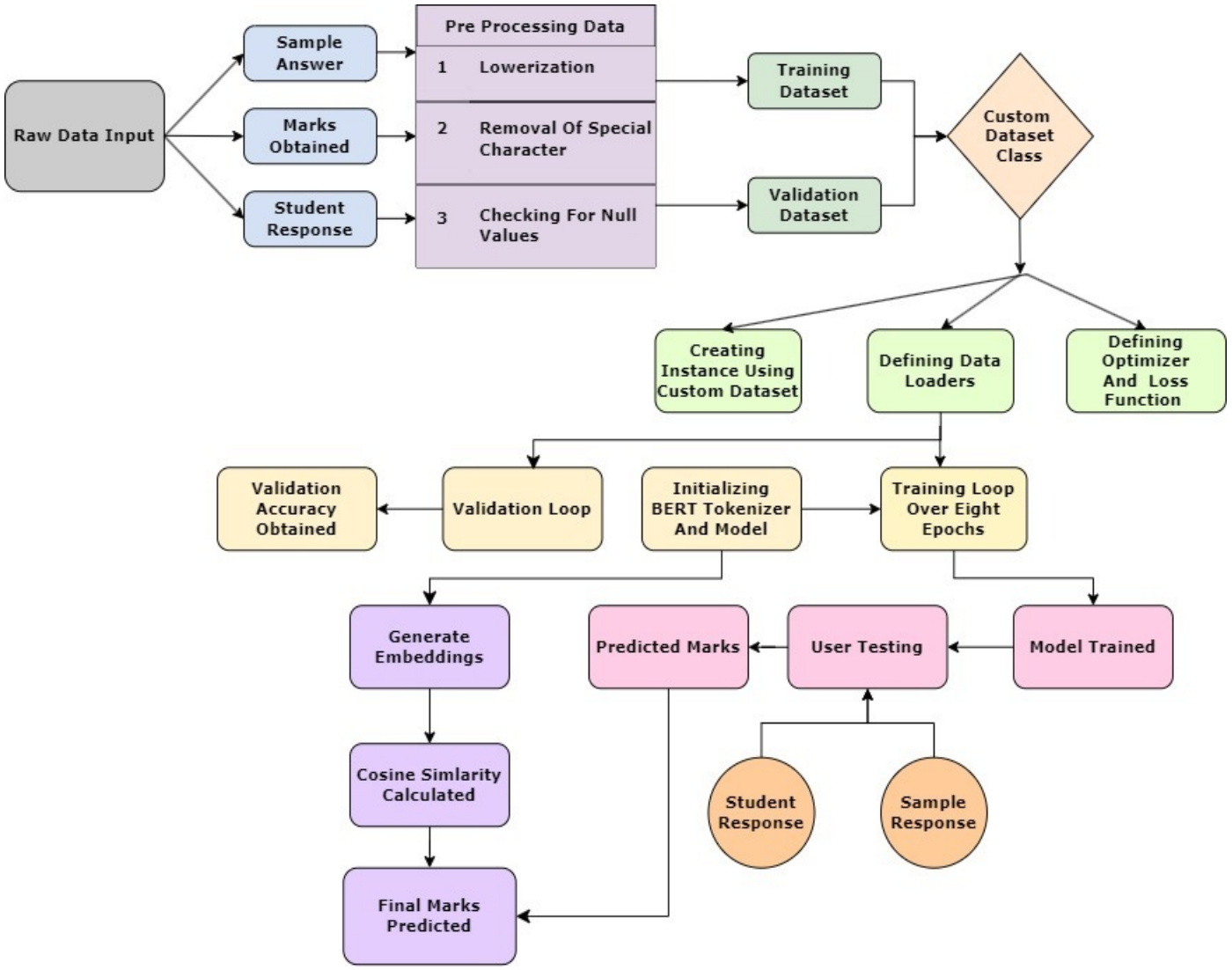


fig 3.3 SCATTER PLOT OF MARKS WITH MAPPED LABEL

Chapter-4 Modelling and Implementation Details

fig 4.1 MODEL PROPOSED SYSTEM FLOW CHART



4.1 Implementation details and issues

Overview:

The implementation of the grading system is centered around a combination of BERT-based sequence classification and a RandomForestClassifier. The goal is to predict grades for student responses based on sample answers and processed responses. The implementation involves data preprocessing, model training, validation, and result evaluation.

Data Preprocessing:

The initial step includes loading the dataset from an Excel file and mapping continuous marks to discrete labels. The dataset is split into training and validation sets. The CustomDataset class is designed to handle data, utilizing the BERT tokenizer for encoding textual information.

Model Architecture:

The model architecture consists of a BERTForSequenceClassification model initialized with pre-trained weights. The training process involves optimizing model parameters using the AdamW optimizer and the CrossEntropyLoss function. Data loaders efficiently handle batches of data during both training and validation.

Challenges and Issues:

Computational Resources:

Training a BERT-based model can be computationally intensive, especially with large datasets and complex architectures. Consideration should be given to available resources, potentially necessitating the use of GPU accelerators.

Data Quality and Size:

The effectiveness of the model is highly dependent on the quality and quantity of the training data. Incomplete or biased datasets may lead to suboptimal performance.

Hyperparameter Tuning:

Fine-tuning hyperparameters, such as learning rates or batch sizes, is crucial for achieving optimal model performance. This process can be time-consuming and may require multiple iterations.

Ensemble Model Integration:

Combining predictions from different models introduces challenges related to model compatibility and integration. Ensuring a seamless ensemble process requires careful consideration.

Evaluation Metrics:
Determining appropriate evaluation metrics is critical. Accuracy alone may not provide a comprehensive view of model performance. Exploring additional metrics such as precision, recall, and F1 score can offer a more nuanced assessment.

4.2 Risk Analysis and Mitigation

Computational Resource Constraints:
Risk: Limited computational resources may hinder the training and evaluation of complex models.
Mitigation: Implementing model parallelism, optimizing code for efficiency, and exploring cloud-based solutions can help mitigate resource constraints.

Data Quality and Bias:
Risk: Poor-quality or biased training data may lead to inaccurate and unfair predictions.
Mitigation: Conducting thorough data preprocessing, addressing biases in the dataset, and employing techniques like data augmentation can enhance data quality.

Hyperparameter Sensitivity:
Risk: The model's performance is sensitive to hyperparameter choices, and suboptimal settings can result in poor predictions.
Mitigation: Conducting systematic hyperparameter tuning, leveraging techniques like grid search or random search, and using cross-validation to evaluate different configurations.

Model Generalization:
Risk: The model may struggle to generalize to unseen data, impacting its real-world applicability.
Mitigation: Incorporating techniques such as transfer learning, fine-tuning on diverse datasets, and exploring domain-specific pre-training can enhance generalization.

Ensemble Model Complexity:
Risk: Integrating multiple models into an ensemble introduces complexity and potential integration challenges.
Mitigation: Ensuring consistent input/output interfaces for models, thorough testing during ensemble construction, and documenting the ensemble process for future maintenance.

Evaluation Metric Selection:
Risk: Relying solely on accuracy as an evaluation metric may not capture the full picture of model performance.
Mitigation: Considering a range of evaluation metrics (precision, recall, F1 score) to provide a comprehensive assessment of the model's strengths and weaknesses.

Future Risks and Mitigations:

Model Drift:
Risk: Over time, the model's performance may degrade as the distribution of input data changes.
Mitigation: Implementing regular model retraining, monitoring data distribution shifts, and updating the model as needed to address drift.

Ethical Considerations:
Risk: The model may inadvertently perpetuate biases present in the training data.
Mitigation: Conducting regular bias audits, incorporating fairness-aware techniques, and involving diverse stakeholders in the model development process.

Regulatory Compliance:
Risk: Changes in regulations related to data privacy or model fairness may impact the system's compliance.
Mitigation: Staying informed about relevant regulations, implementing robust privacy measures, and maintaining transparency in the model's decision-making process.

Chapter-5 Testing

5.1 Testing Plan

Objective:
The testing plan has a two fold objective: to ensure the reliability, accuracy, and robustness of the grading system while also identifying and mitigating potential issues that may arise during its implementation. The plan encompasses various testing phases, each designed to validate different aspects of the system's functionality, from individual components to the entire end-to-end grading process.

Testing Phases:

1. Unit Testing:
In the unit testing phase, individual components of the grading system will undergo scrutiny. This includes testing data preprocessing functions, the functionality of the CustomDataset class methods, model training functions, and the correctness of evaluation metric calculations. By ensuring that each function operates as expected, validating data transformations and encoding, and confirming proper model parameter updates during training, this phase lays the foundation for robust component-level functionality.

2. Integration Testing:
Moving beyond individual components, the integration testing phase focuses on how these components interact with each other. This includes testing the seamless flow of data between preprocessing and model training, ensuring the integration of BERT-based, and validating the integration of these diverse models in the ensemble. By confirming compatibility between different model outputs and the successful interaction between system components, this phase ensures the harmonious functioning of the entire system.

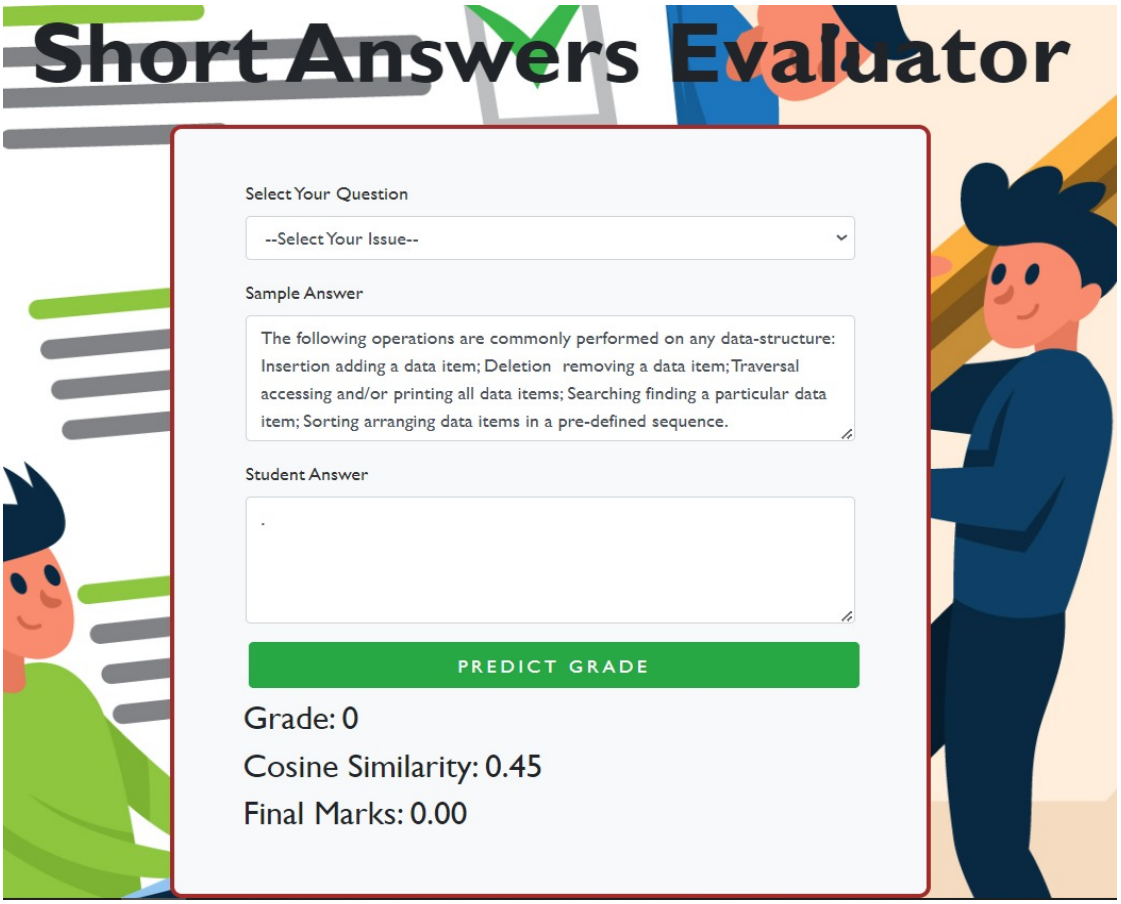
3. System Testing:
In the system testing phase, the entire grading system undergoes evaluation in real-world scenarios. This includes an end-to-end assessment of the grading process from data loading to result generation. The system's capability to handle varying input sizes and formats is tested, and the ensemble model's predictions are thoroughly evaluated. By validating overall system functionality, ensuring adaptability

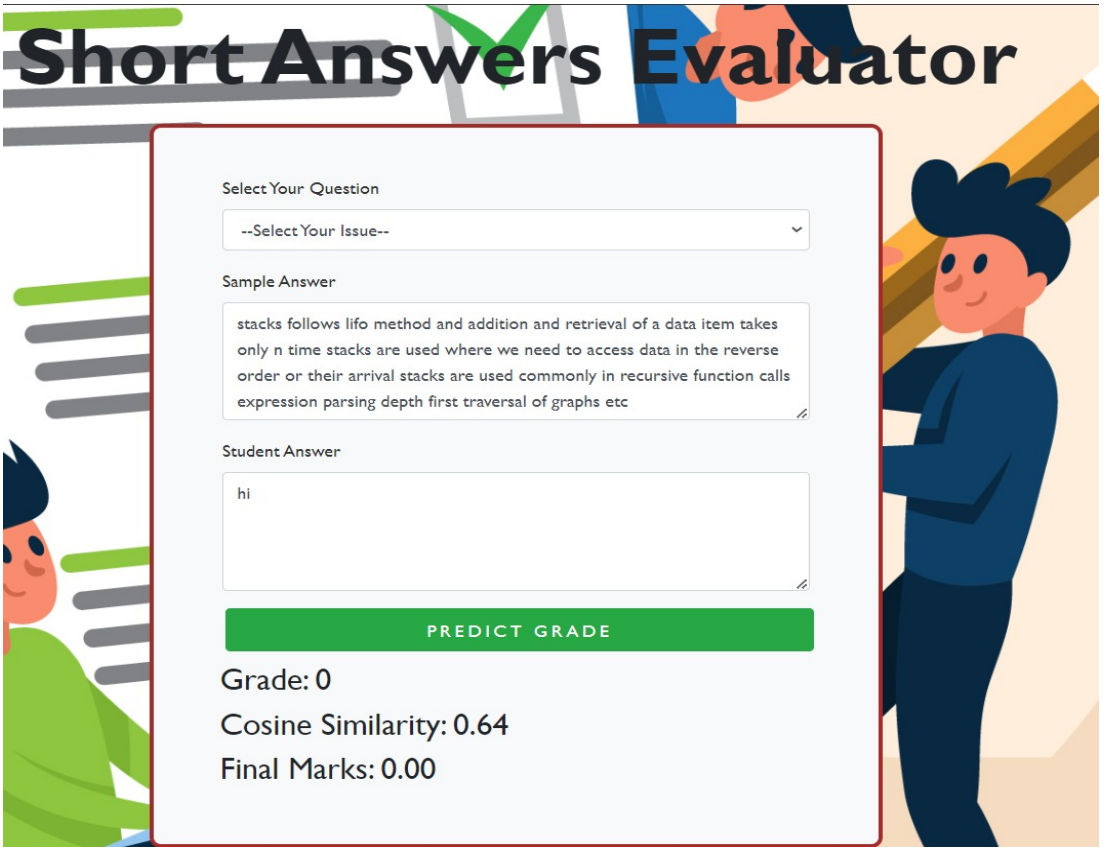
to diverse inputs, and confirming the accuracy and consistency of results, this phase provides a comprehensive assessment of the system's performance.

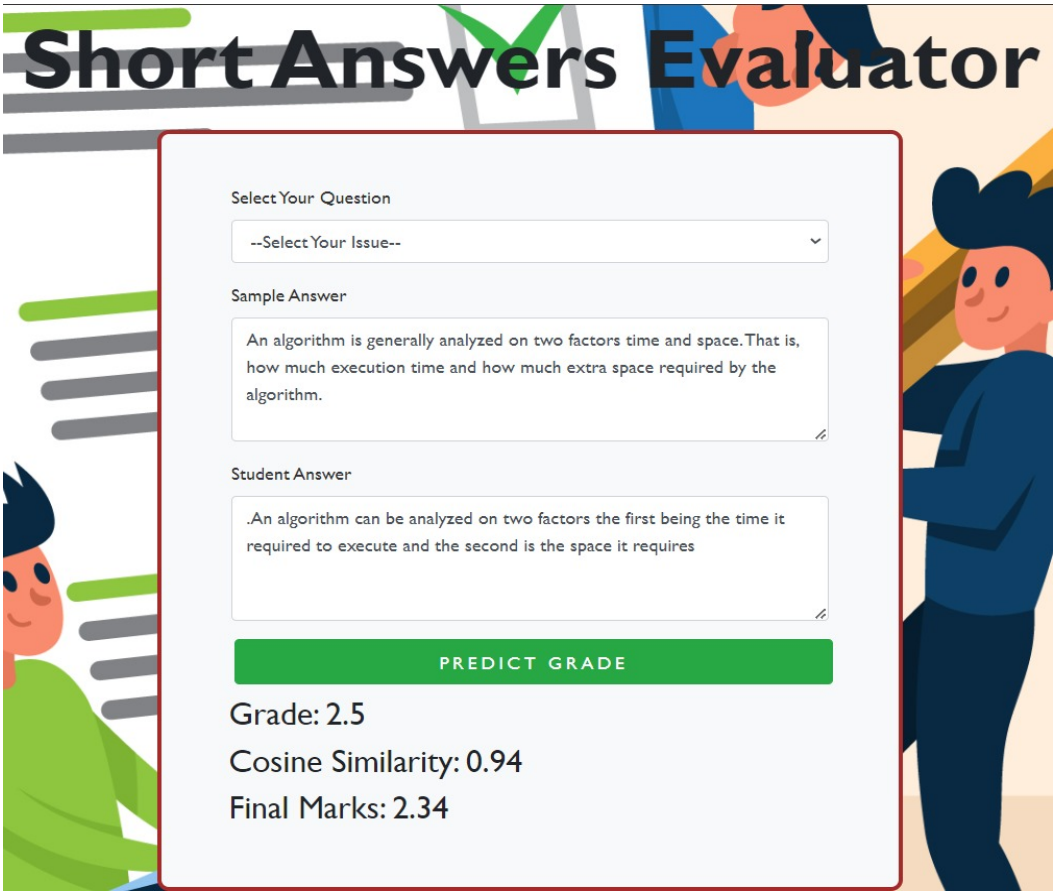
4. Performance Testing:
Performance testing focuses on assessing the system's efficiency. Metrics such as training time for BERT and models, inference time for individual models and the ensemble, and memory usage during training and inference are measured. The goal is to ensure that the models meet performance expectations, identify potential bottlenecks, and optimize the system for efficient operation.

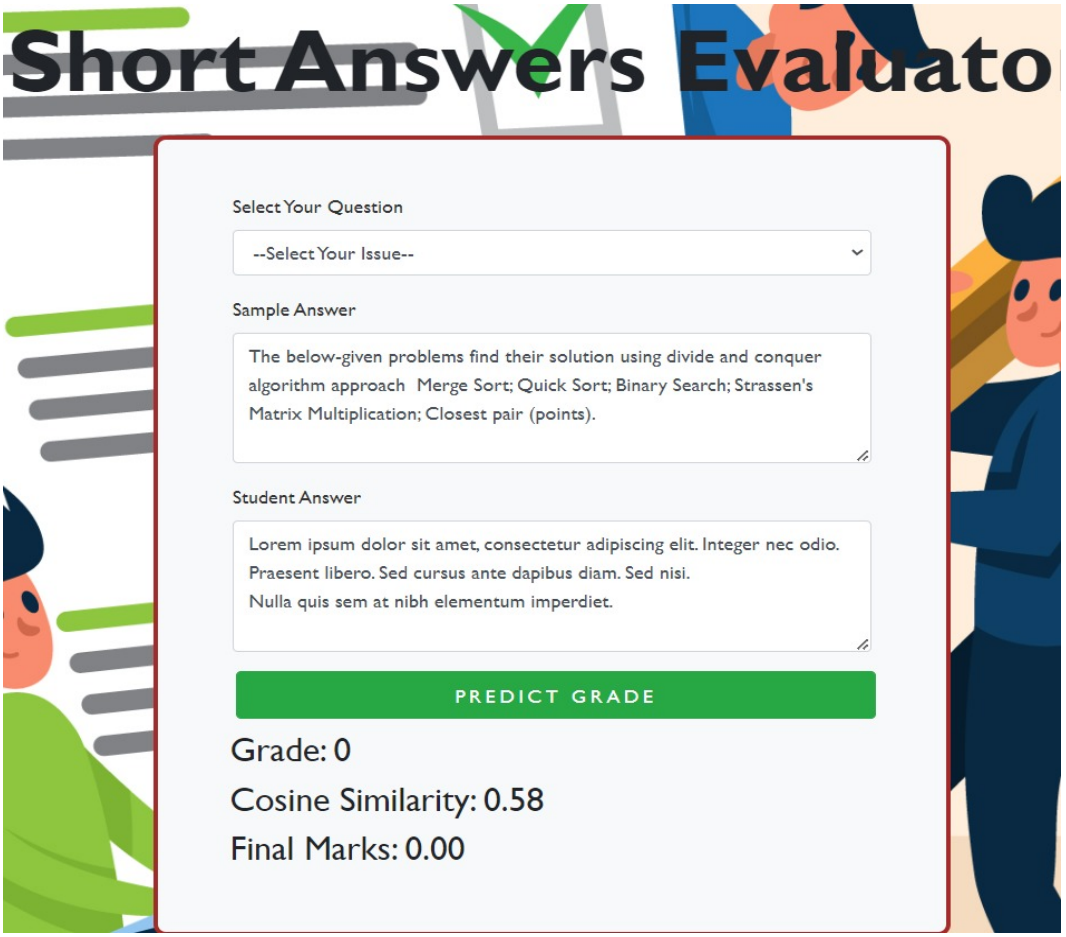
5. User Acceptance Testing (UAT):
The final phase involves user acceptance testing to evaluate the system's usability and effectiveness from an end-user perspective. Real-world student responses are used to assess the system's predictions, and user feedback is gathered to validate the system's usability. By obtaining user input and verifying that predicted grades align with user expectations, this phase ensures that the grading system meets user requirement

5.2 Test Cases


Short Answers Evaluator
Select Your Question
--Select Your Issue--
Sample Answer
The following operations are commonly performed on any data-structure: Insertion adding a data item; Deletion removing a data item; Traversal accessing and/or printing all data items; Searching finding a particular data item; Sorting arranging data items in a pre-defined sequence.
Student Answer
.
PREDICT GRADE
Grade: 0
Cosine Similarity: 0.45
Final Marks: 0.00


Short Answers Evaluator
Select Your Question
--Select Your Issue--
Sample Answer
stacks follows lifo method and addition and retrieval of a data item takes only n time stacks are used where we need to access data in the reverse order or their arrival stacks are used commonly in recursive function calls expression parsing depth first traversal of graphs etc
Student Answer
hi
PREDICT GRADE
Grade: 0
Cosine Similarity: 0.64
Final Marks: 0.00


Short Answers Evaluator
Select Your Question
--Select Your Issue--
Sample Answer
An algorithm is generally analyzed on two factors time and space. That is, how much execution time and how much extra space required by the algorithm.
Student Answer
.An algorithm can be analyzed on two factors the first being the time it required to execute and the second is the space it requires
PREDICT GRADE
Grade: 2.5
Cosine Similarity: 0.94
Final Marks: 2.34


Short Answers Evaluato
Select Your Question
--Select Your Issue--
Sample Answer
The below-given problems find their solution using divide and conquer algorithm approach Merge Sort; Quick Sort; Binary Search; Strassen's Matrix Multiplication; Closest pair (points).
Student Answer
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Integer nec odio. Praesent libero. Sed cursus ante dapibus diam. Sed nisi.
Nulla quis sem at nibh elementum imperdiet.
PREDICT GRADE
Grade: 0
Cosine Similarity: 0.58
Final Marks: 0.00


Short Answers Evaluato
Select Your Question
--Select Your Issue--
Sample Answer
A circular linked list is a special type of linked list that supports traversing from the end of the list to the beginning by making the last node point back to the head of the list.
Student Answer
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Integer nec odio. Praesent libero. Sed cursus ante dapibus diam. Sed nisi. Nulla quis sem at nibh elementum imperdiet.
PREDICT GRADE
Grade: 2.5
Cosine Similarity: 0.54
Final Marks: 1.34

5.3 Limitations of the solution

While our grading system based on BERT-based models and ensemble methods offers a robust solution, it is essential to acknowledge its limitations. Understanding these limitations is crucial for managing expectations and guiding future improvements.

Data Dependence:

Our model's performance heavily relies on the quality and diversity of the training data. Limitations in the dataset, such as biases or lack of representation, can impact the model's generalization to real-world scenarios.

Computational Resources:

Training and utilizing BERT-based models require significant computational resources, limiting the accessibility of the solution. Not all educational institutions or users may have access to high-performance computing environments.

Interpretability:

BERT-based models are complex, and their decision-making processes are often considered black-box. Understanding the rationale behind specific predictions might be challenging, especially in educational contexts where transparency and interpretability are crucial.

Domain Specificity:

The grading system might be more effective in certain domains or subjects than others. Adapting the system to diverse academic disciplines may require additional fine-tuning or domain-specific models.

Dynamic Grading Criteria:

The system assumes a static mapping of labels to grades. In dynamic educational settings where grading criteria evolve, the model may need frequent updates to align with changing standards.

Limited Explainability:

While efforts have been made to interpret the model's predictions through ensemble methods, the overall explainability of the grading decisions may not satisfy all stakeholders, such as educators, students, or administrators.

Overfitting Risks:

In complex models like BERT, there is a risk of overfitting, especially if the model is exposed to a limited set of responses. This could result in the model performing well on training data but struggling with unseen examples.

Resource Intensiveness during Inference:

Deploying the grading system in real-time environments might face challenges due to the resource-intensive nature of BERT models during inference, impacting the system's responsiveness.

Ensemble Complexity:

While ensemble methods aim to enhance predictive performance, managing and maintaining the ensemble's complexity can be challenging. The benefits of ensemble learning come with increased computational costs and potential challenges in real-world deployment.

Scalability:

As the grading system becomes popular and usage scales, issues related to scalability might arise. Handling a large number of concurrent grading requests may necessitate infrastructure adjustment.

Chapter-6 Findings, Conclusions and Future Work

6.1 Findings

The findings related to the manual system of answer checking underscore a multitude of challenges and inefficiencies that warrant a transition towards a modern system leveraging Natural Language Processing (NLP). The conventional manual approach to evaluating subjective answers, often reliant on human examiners, proves to be time-consuming, susceptible to errors, and inherently subjective.

In the manual system, examiners face the arduous task of individually assessing each answer, often leading to delays in result dissemination. The time-consuming nature of this process not only hinders the timely feedback crucial for student learning but also poses logistical challenges, especially in large-scale examinations where a significant volume of answer scripts requires evaluation.

Moreover, the manual system is prone to subjective biases, influenced by individual examiner perspectives, preferences, and potential fatigue. Consistency in grading across different examiners becomes a significant concern, impacting the fairness of evaluations. Additionally, the manual approach struggles to handle the nuances of natural language, often leading to variations in grading based on subjective interpretations.

Errors and inaccuracies in grading further contribute to the limitations of the manual system. Human examiners may inadvertently introduce inconsistencies or overlook subtle nuances, impacting the overall reliability of the evaluation process. The potential for misinterpretation of handwriting or subjective answers adds an additional layer of complexity to the manual system.

Recognizing the inadequacies of the manual approach, the imperative for an upgrade to a modern system using NLP becomes evident. Natural Language Processing, a field of artificial intelligence, holds the promise of addressing these challenges and revolutionizing the subjective answer evaluation process.

By implementing NLP techniques, the modern system can swiftly analyze and process vast amounts of textual data, significantly reducing the time required for evaluation. The efficiency gains are particularly pronounced in the context of large-scale examinations, where NLP can expedite the grading process without compromising accuracy.

Furthermore, NLP introduces objectivity and consistency in answer evaluations. The system adheres to predefined criteria, eliminating the influence of subjective biases that may impact human examiners. This ensures a fair and uniform assessment across all answer scripts, enhancing the overall credibility of the evaluation process.

The capabilities of NLP extend to handling the intricacies of natural language. Algorithms can be designed to recognize synonyms, contextual nuances, and subtle variations in language, mitigating the limitations associated with the manual interpretation of subjective answers.

In conclusion, the findings advocate for a paradigm shift from the manual system of answer checking to a modernized approach utilizing NLP. This transition promises to streamline the evaluation process, enhance efficiency, eliminate biases, and improve the overall accuracy and reliability of subjective answer assessments. The integration of NLP technologies aligns with the imperative for educational institutions to embrace innovative solutions that not only address existing challenges but also pave the way for a more robust and equitable evaluation framework.

6.2 Conclusion

In this minor project, a BERT-based model was successfully implemented for grading student responses based on a provided sample answer. The dataset underwent preprocessing and transformation into a format suitable for training the model. The Hugging Face Transformers library was utilized for working with the BERT model, and the PyTorch framework was employed for training and evaluation. The model underwent training on a labeled dataset, and its performance was evaluated on a validation set, thereby showcasing the effectiveness of the implemented solution.

The convergence of the model over multiple epochs, optimization of the cross-entropy loss with the AdamW optimizer, and the subsequent saving of the trained model and tokenizer for potential future use were demonstrated in the training loop. Following this, a validation loop was executed to assess the accuracy of the model on the validation set, revealing the overall effectiveness of the model in predicting student grades.

For the practical application of the model, a function for predicting grades based on sample and student responses was implemented. The model's predictions were applied to a subset of the validation dataset, generating a DataFrame that included actual and predicted grades, along with the corresponding labels. This DataFrame facilitated a detailed examination of the model's performance on individual responses.

Furthermore, a comprehensive analysis of the model's predictions was conducted through the construction and visualization of a confusion matrix. This facilitated a deeper understanding of the strengths and weaknesses of the model in classifying different grades.

fig 6.1 PLOT OF TRAINING LOSS VS EPOCH

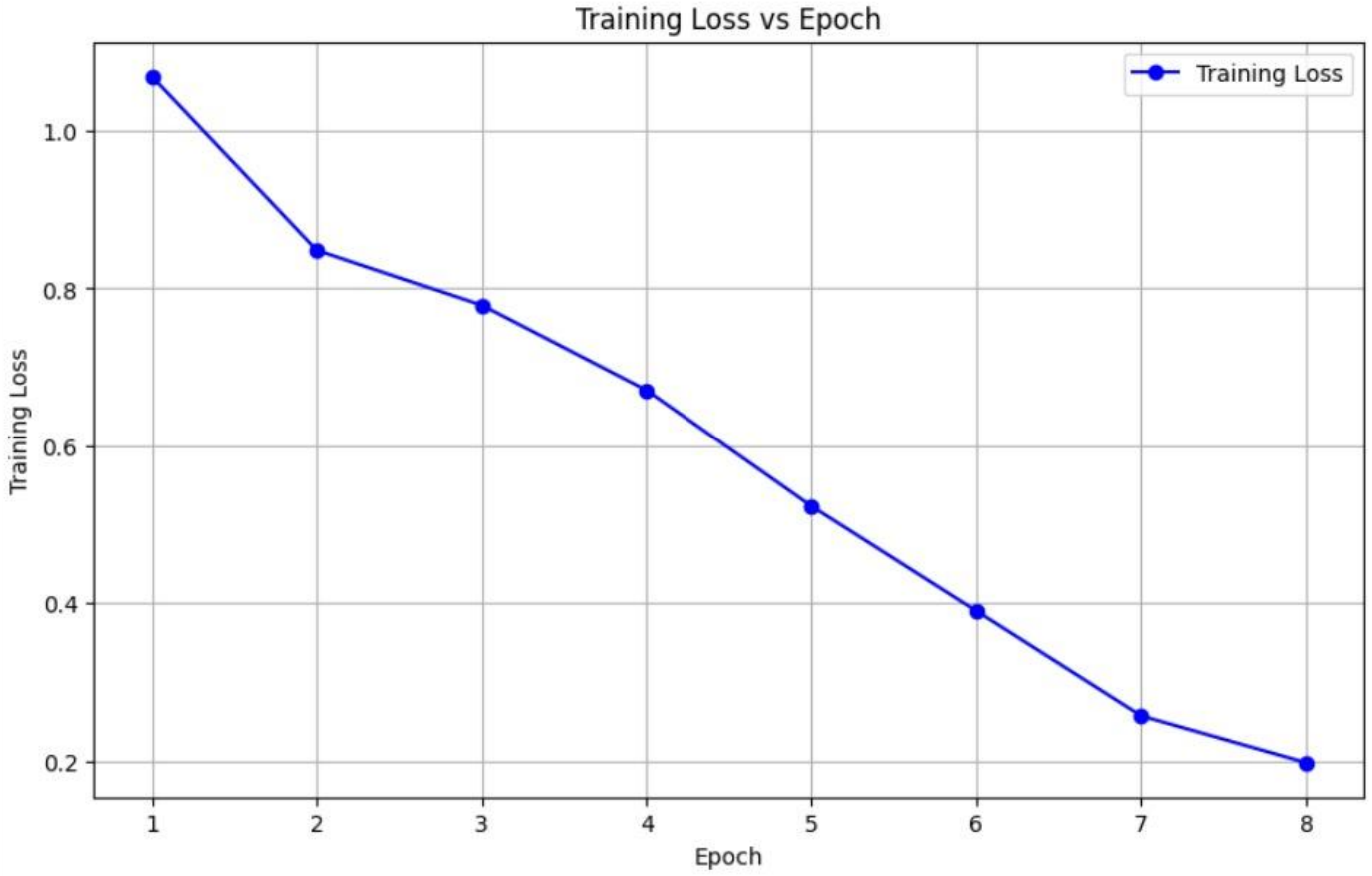


fig 6.2 MODEL ACCURACY ON VALIDATION DATASET

```
# Calculate accuracy on the validation set
val_accuracy = accuracy_score(all_labels, all_predictions)
print(f'Validation Accuracy: {val_accuracy:.4f}')
```

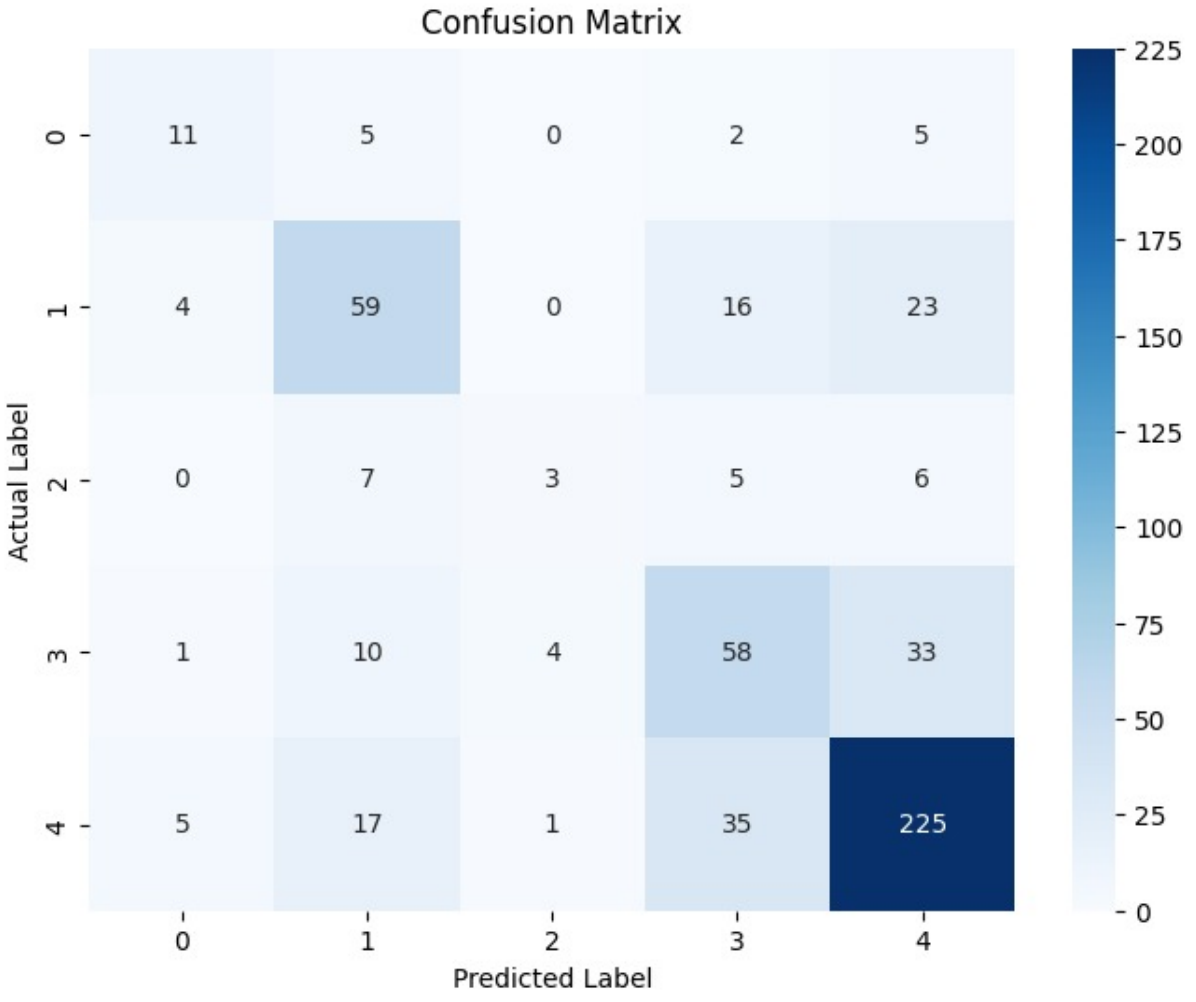
```
Validation Accuracy: 0.6879
```

fig 6.3 F1 SCORE FOR EACH PREDICTED LABEL

```
Class 0: Precision=0.8571, Recall=0.5217, F1 Score=0.6486
Class 1: Precision=0.0000, Recall=0.0000, F1 Score=0.0000
Class 2: Precision=0.6392, Recall=0.6078, F1 Score=0.6231
Class 3: Precision=0.4000, Recall=0.2857, F1 Score=0.3333
Class 4: Precision=0.4651, Recall=0.5660, F1 Score=0.5106
Class 5: Precision=0.8143, Recall=0.8057, F1 Score=0.8099

Macro Average: Precision=0.6351, Recall=0.5574, F1 Score=0.5851
Micro Average: Precision=0.6879, Recall=0.6879, F1 Score=0.6879
```

fig 6.4 CONFUSION MATRIX BETWEEN ACTUAL LABEL AND PREDICTED LABEL



6.3 Future Work

In the envisioned future developments, the integration of our model will extend beyond the current scope to incorporate semantic graphs and knowledge graphs. Shifting the model's training paradigm from sample responses to a broader dataset is a priority. This transition aims to enhance the model's understanding by exposing it to diverse and extensive educational data, fostering improved generalization capabilities.

Furthermore, the pursuit of advancements will involve meticulous fine-tuning of hyperparameters and exploration of alternative model architectures. Techniques like hyperparameter optimization and experimentation with ensemble methods will be pivotal in refining the model's performance and bolstering its overall robustness.

To enhance interpretability, future iterations may consider incorporating attention mechanisms or leveraging gradient-based methods. These enhancements will shed light on the specific aspects of input data influencing the model's predictions. This interpretability facet holds significance, especially in educational contexts, as it provides valuable insights for delivering constructive feedback to students.

Evaluation on real-world student responses will be a crucial step forward, facilitating the implementation of iterative feedback loops. Collaborating with educators and domain experts will be instrumental in fine-tuning grading criteria to ensure alignment with educational standards.

In conclusion, while the current project establishes a strong foundation for automatic grading using BERT, the envisioned future developments, such as semantic and knowledge graph integration and training on comprehensive datasets, aim to elevate the model's capabilities. The iterative nature of these advancements underscores a commitment to continuous improvement, positioning this project as a transformative step toward more sophisticated and effective automatic grading systems.

References:

1. https://ieeexplore.ieee.org/document/10064615
2. NLP-based Automatic Answer Evaluation | IEEE Conference Publication | IEEE Xplore
3. https://www.mililink.com/upload/article/945421994aams_vol_2011_september_2021_a34_p2749-2765_b._madhavi_desai_et_al..pdf
4. https://ieeexplore.ieee.org/document/9627669
5. Subjective Answers Evaluation Using Machine Learning and Natural Language Processing | IEEE Journals & Magazine | IEEE Xplore
6. Word embeddings in NLP: A Complete Guide (turing.com)