IMPLEMENTATION OF LARGE LANGUAGE MODEL FOR AUTOMATIC ANSWER SCRIPT EVALUATION

BY

MIR MYNUL AHASAN MIM
ID: 201-15-13953
AND

SADIKUL ISLAM
ID: 201-15-13983

FINAL YEAR DESIGN PROJECT REPORT

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Dr. FIZAR AHMED
Associate Professor
Department of Computer Science and Engineering
Daffodil International University

Co-Supervised By

SHARMIN AKTER RIMA
Lecturer (Senior Scale)
Department of Computer Science and Engineering
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

July 2024

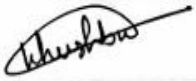
APPROVAL

This Project titled "Implementation of Large Language Model for Automatic Answer Script Evaluation", submitted by **[Mir Mynul Ahasan Mim]** and **[Sadikul Islam]** to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 15-07-2024.

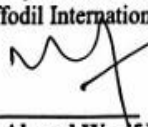
BOARD OF EXAMINERS

Dr. Md. Ismail Jabiullah (MIJ) — **Board Chairman**
Professor
Department of CSE
Faculty of Science & Information Technology
Daffodil International University

Mr. Md Assaduzzaman (MA) — **Internal Examiner 1**
Lecturer (Senior Scale)
Department of CSE
Faculty of Science & Information Technology
Daffodil International University



Ms. Sharun Akter Khushbu (SAK) — **Internal Examiner 2**
Lecturer (Senior Scale)
Department of CSE
Faculty of Science & Information Technology
Daffodil International University



Dr. Ahmed Wasif Reza (DWR) — **External Examiner**
Professor
Department of Computer Science & Engineering
East West University

©Daffodil International University, Bangladesh

©Daffodil International University, Bangladesh

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Dr. Fizar Ahmed, Associate Professor, Department of Computer Science and Engineering, Daffodil International University**. We also declare that neither this project nor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

Dr. Fizar Ahmed
Associate Professor
Department of CSE
Daffodil International University

Co-Supervised by:

Sharmin Akter Rima
Lecturer (Senior Scale)
Department of CSE
Daffodil International University

Submitted by:

(Mir Mynul Ahasan Mim)
ID: 201-15-13953
Department of CSE
Daffodil International University

(Sadikul Islam)
ID: 201-15-13983
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to almighty for His divine blessing making it possible for us to complete the final year project/internship successfully.

We are grateful and wish our profound indebtedness to **Dr. Fizar Ahmed**, **Associate Professor**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "*Natural Language Processing*" to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts, and correcting them at all stages have made it possible to complete this project.

We would like to express our heartiest gratitude to the **Head of the Department of CSE**, for his kind help in finishing our project and also to other faculty members and the staff of the Department of CSE, Daffodil International University.

We would like to thank our entire course mate in Daffodil International University, who took part in this discussion while completing the course work.

Finally, we must acknowledge with due respect the constant support and patience of our parents.

ABSTRACT

This paper explores the implementation of large language models (LLMs) for automatic answer script evaluation, examining their potential to revolutionize educational assessment. LLMs offer significant advantages, including increased efficiency, consistency, and scalability in grading processes, as well as cost savings for educational institutions. However, their deployment also presents challenges, such as high energy consumption, potential biases, privacy concerns, and the socio-economic impact on employment within the educational sector. This study addresses these issues by proposing a comprehensive sustainability plan that includes optimizing energy efficiency, utilizing renewable energy sources, enhancing data protection measures, and ensuring transparency and accountability in automated evaluations. Additionally, strategies for balancing automation with human oversight and upskilling educators are discussed. The findings highlight the need for ongoing research to refine LLM applications, ensuring their ethical, sustainable, and effective integration into educational systems. The implications for further study emphasize the importance of continuous improvement in energy use, bias mitigation, data privacy, and the broader impact on educational practices and employment.

TABLE OF CONTENTS

CONTENTS	PAGE
Board of examiners	ii
Declaration	iii
Acknowledgements	iv
Abstract	v
Table of contents	vi-viii
List of Figures	ix-x
List of Tables	xi
CHAPTER	
CHAPTER 1: INTRODUCTION	**12-14**
1.1 Overview	12
1.2 Background and Present State	12
1.3 Problem Statement	12
1.4 Objective	13
1.5 Scope and Limitation	13
1.6 Report Organization	13
1.7 Summery	14

CHAPTER 2: LITERATURE REVIEW **15-19**

2.1 Overview 15

2.2 Related Works 15-17

2.3 Comparative between existing works 17-19

2.4 Open Issues 19

2.5 Summery 19

CHAPTER 3: METHODOLOGY / REQUIREMENT ANALYSIS & DESIGN SPECIFICATION **20-25**

3.1 Overview 20

3.2 Proposed Methodology / System Design 20-25

3.3 Hardware / Software Requirement 24-25

3.4 Project Management and Financial Analysis 25

3.5 Summery 25

CHAPTER 4: IMPLEMENTATION **26-29**

4.1 Overview 26-27

4.2 Train Model / Prototype Design 27-28

4.3 System Testing / Model Evaluation 28-29

4.4 Summary 29

CHAPTER 5: RESULT AND ANALYSIS **30-39**

5.1 Overview 30

5.2 Experimental / Simulation Result 30-37

5.3 Performance / Comparative Analysis 38

5.4 Summery 38-39

CHAPTER 6: IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY **39-44**

6.1 Impact on Life **39-41**

6.2 Impact on Society & Environment **41-42**

6.3 Ethical Aspects **42**

6.4 Sustainability Plan **43**

6.5 Summary **43-44**

CHAPTER 7: CONCLUSION AND FUTURE WORK

7.1 Conclusion **44-45**

7.2 Further Suggested Works **45**

7.3 Limitations / Conflicts of interest **45-46**

REFERENCES **47**

APPENDIX **48-43**

LIST OF FIGURES

FIGURES	PAGE NO
Figure 3.2.1: Proposed Methodology	20
Figure 3.2.2: Datasets	22
Figure 3.2.3: Preprocessing	23
Figure 3.2.4: Schematic representation of DistilBERT	23
Figure 3.2.5: Schematic Representation of BERT	24
Figure 3.2.2: Schematic representation of RoBERTa	24
Figure 5.2.4: Training and validation accuracy and loss over the epochs (RoBERTa).	32
Figure 5.2.5: ROC Curve for RoBERTa	33
Figure 5.2.6: Training and validation accuracy and loss over the epochs (DistilBERT).	33-34
Figure 5.2.7: ROC Curve for DistilBERT	34
Figure 5.2.8: Training and validation accuracy and loss over the epochs (BERT).	35
Figure 5.2.9: ROC Curve for BERT	36
Figure 5.2.10: Confusion Matrix for RoBERTa	36
Figure 5.2.11: Confusion Matrix for DistilBERT	37
Figure 5.2.12: Confusion Matrix for BERT	37

LIST OF TABLES

TABLES	PAGE NO
TABLE 2.3.1: Comparative Analysis Table	17-19
TABLE 3.4.1: Estimated cost of our project	25
TABLE 5.2.1: Precision, Recall, F1-Score, and Support (n) for DistilBERT	30
TABLE 5.2.2: Precision, Recall, F1-Score, and Support (n) for RoBERTa	31
TABLE 5.2.3: Precision, Recall, F1-Score, and Support (n) for BERT	31
TABLE 5.3.1: Accuracy table for different model	38

CHAPTER 1
Introduction

1.1 Overview

The advent of artificial intelligence (AI) has revolutionized numerous fields, including education. One significant application is the automatic evaluation of answer scripts, which leverages large language models (LLMs) to enhance the efficiency and consistency of grading processes. This thesis explores the implementation of LLMs for automatic answer script evaluation, aiming to reduce the workload on educators and provide timely feedback to students. The study delves into the mechanisms of these models, their training processes, and their performance in real-world scenarios, emphasizing their potential to transform traditional educational assessments.

1.2 Background and Present State

Automatic answer script evaluation has been an area of active research, particularly with the increasing sophistication of natural language processing (NLP) techniques. Traditional methods relied on keyword matching and rule-based systems, which often lacked the nuanced understanding required for subjective assessments. With the introduction of LLMs such as BERT, GPT, and their variants, there has been a significant improvement in the ability to comprehend and evaluate complex text. Presently, these models are trained on extensive datasets and fine-tuned for specific tasks, demonstrating remarkable accuracy and consistency in grading. However, the adoption of these technologies in educational institutions is still in its nascent stages, with ongoing research focused on improving their reliability and scalability.

1.3 Problem Statement

Despite the promising capabilities of LLMs, several challenges hinder their widespread implementation in automatic answer script evaluation. These include the need for large, annotated datasets for training, potential biases in model predictions, and the lack of transparency in decision-making processes. Additionally, the integration of these models into existing educational frameworks requires careful consideration of ethical and privacy concerns. This thesis addresses these issues by developing a robust LLM-based system for answer script evaluation, evaluating its performance, and proposing solutions to overcome identified challenges.

©*Daffodil International University*

1.4 Objective

The primary objective of this thesis is to design and implement a system for automatic answer script evaluation using LLMs. The study aims to achieve the following specific goals:

1. Develop a model capable of understanding and grading complex answers with high accuracy.
2. Assess the performance of the model against traditional grading methods.
3. Identify and mitigate potential biases in the model's predictions.
4. Ensure the system's compliance with ethical standards and data privacy regulations.
5. Provide recommendations for integrating the system into educational institutions.

1.5 Scope and Limitation

The scope of this thesis encompasses the development and evaluation of an LLM-based answer script evaluation system. The research focuses on the model's ability to grade subjective and objective questions across various subjects. However, the study is limited by the availability of large, annotated datasets, which are essential for training and fine-tuning the model. Additionally, the system's performance may vary based on the complexity of the subject matter and the diversity of the answer scripts. The ethical and privacy considerations associated with using student data also pose limitations that need to be addressed.

1.6 Report Organization

The thesis is organized into several chapters, each addressing a specific aspect of the research. Chapter 1 provides an introduction to the study, including the overview, background, problem statement, objectives, scope, and limitations. Chapter 2 reviews the existing literature on automatic answer script evaluation and the use of LLMs in education. Chapter 3 describes the methodology used to develop and evaluate the proposed system. Chapter 4 presents the results and discusses the findings in the context of the research objectives. Chapter 5 concludes the thesis with a summary of the key contributions, implications for practice, and suggestions for future research.

1.7 Summary

This introductory chapter has outlined the motivation behind the study, highlighting the potential of LLMs to revolutionize answer script evaluation in educational settings. The background section

provided context to the present state of research in this area, while the problem statement identified the key challenges that this thesis aims to address. The objectives, scope, and limitations set the boundaries for the study, ensuring a focused and manageable research process. The organization of the report was also detailed, providing a roadmap for the subsequent chapters. The following sections will delve deeper into the methodologies employed and the results obtained, leading to a comprehensive understanding of the potential and limitations of using LLMs for automatic answer script evaluation. It introduces large language models (LLMs) like DistilBERT, BERT, and RoBERTa as promising solutions for automating the evaluation process. These models offer the potential for faster, more consistent, and fairer assessments, benefiting both educators and students. The introduction sets the context for exploring the implementation of LLMs in automatic answer script evaluation, emphasizing their potential to transform educational assessments.

CHAPTER 2
Background

2.1 Overview

The literature review provides a comprehensive examination of existing research related to automatic answer script evaluation using large language models (LLMs). This chapter synthesizes key findings from previous studies, highlighting the evolution of techniques and the state-of-the-art methods in the field. It aims to contextualize the current study within the broader landscape of educational technology and natural language processing, demonstrating the progression from traditional grading systems to advanced AI-driven approaches.

2.2 Related Works

Numerous studies have explored the use of artificial intelligence in automating the evaluation of academic answer scripts. Early approaches primarily relied on keyword matching and rule-based systems, which were limited in their ability to handle nuanced and context-dependent answers. The advent of machine learning and natural language processing (NLP) introduced more sophisticated techniques, such as latent semantic analysis and support vector machines. Recently, the development of LLMs like BERT, GPT, and RoBERTa has marked a significant advancement in the field. These models have demonstrated superior performance in understanding and evaluating complex textual responses due to their deep contextual comprehension and extensive pre-training on large datasets.

The field of automatic answer evaluation has seen significant advancements in recent years, largely driven by the application of machine learning and deep learning techniques. Various research studies have explored different methodologies to improve the accuracy and efficiency of automated grading systems.

One prominent study by Nandita Bharambe et al. (2021) aimed to develop an automated descriptive answer evaluation system using machine learning. This system evaluates answers based on matched keywords and the minimum length specified by the moderator. It employs Optical Character Recognition (OCR) to convert handwritten scanned answer sheets into machine-readable text and utilizes an artificial neural network (ANN) with a back-propagation algorithm to grade the answers. The system's effectiveness is measured by matching extracted

keywords with those provided by the moderator and considering the length of the answers.

In a similar vein, the research conducted by M. Syamala Devi and Himani Mittal (2013) applied Latent Semantic Analysis (LSA) for the evaluation of subjective answers. Their study developed a prototype using MATLAB and Java to evaluate technical answers by clustering text into groups based on similarity. The LSA technique proved effective in comparing student answers with model answers, yielding satisfactory results that were comparable to human evaluation.

The work by David Jackson and Michelle Usher (1997) on the ASSYST system showcased the potential of automating the grading of student programs. ASSYST offers a graphical interface for directing the grading process and applies various metrics such as correctness, efficiency, complexity, style, and test data adequacy to evaluate student programs. The system has been positively received for its ability to provide consistent and thorough assessments.

Another approach is highlighted in the study by R. P.-Rodriguez et al. (2022), which developed a prediction model for pediatric radiographic pneumonia using clinical variables, the mNUTRIC score, and pneumonia prediction rules. This model demonstrated high accuracy in predicting clinical outcomes, suggesting the utility of machine learning models in medical diagnostics and prognosis.

Similarly, J. D. Bodapati et al. (2022) developed a deep neural network model to assess juvenile pneumonia in chest radiographs. Their model outperformed manual examination in various performance metrics, simplifying pneumonia detection for both professionals and novices. Their accuracy was approximately 53%.

E. ERDEM et al. (2021) compared pre-trained and proposed deep learning models for pneumonia detection. Their new deep neural network achieved higher recall and F1-scores compared to established models like VGG16 and VGG19, highlighting the advancements in deep learning architectures for medical imaging. Their accuracy was approximately 56%.

Li Bin, et al (2008) The K-Nearest Neighbor (KNN) algorithm for text categorization is applied to CET4 essays. They tried to research each essay is represented by the Vector Space Model (VSM). After removing the stop words, we chose the words, phrases and arguments as features of the essays, and the value of each vector is expressed by the term frequency and inversed document frequency (TF-IDF) weight. The TF and information fain (IG) methods are used to select features by predetermined thresholds. We calculated the similarity of essays with cosine in the KNN algorithm. Experiments on CET4 essays in the Chinese Learner English Corpus (CLEC) show accuracy above 76% is achieved.

Overall, these studies illustrate the diverse applications of machine learning and deep learning in automatic answer evaluation, medical diagnostics, and program grading. The advancements in these fields continue to enhance the accuracy, efficiency, and reliability of automated evaluation systems, providing valuable tools for educators and healthcare professionals alike.

2.3 Comparison between Existing Works

Comparative studies between traditional methods and LLM-based approaches reveal distinct advantages of the latter. Traditional methods often struggle with the variability and complexity of student answers, leading to inconsistent and sometimes inaccurate grading. In contrast, LLMs leverage deep learning techniques to capture intricate patterns and meanings within text, resulting in more reliable and consistent evaluations. Various benchmarks and experiments have shown that models like BERT and RoBERTa outperform earlier NLP methods in terms of accuracy and robustness. However, these advanced models also require significant computational resources and large annotated datasets for effective training and fine-tuning.

TABLE 2.3.1: Comparative Analysis Table

Study	Model	Accuracy
Parag Guruji Mrunal Pagnis et al. (2015)	Generalized Latent Semantic Analysis (GLSA) was the specific algorithm.	Synonyms enhance accuracy in grading subjective answers. Synonym sets improve system performance without significant resource consumption.
M Syamala Devi et al. (2013)	Latent Semantic Analysis (LSA) algorithm was used for evaluation. LSA algorithm clusters text based on similarity for evaluation purposes. LSA algorithm was implemented using Java and MatLab platforms. LSA technique was applied to evaluate technical answers effectively.	GLSA technique achieved 89% accuracy, close to human grading. LSA yielded slightly more accurate grading than PLSA and LDA.
Sheeba Praveen (2014)	Shift reduce style parser based on Adwait Ratnaparki's thesis. Model trained on English data from Wall Street Journal.	POS tagger model achieved over 96% accuracy on unseen data.

Ch Kumar et al. (2012)	FCA, VSM, and LSI were used for information retrieval analysis. FCA and VSM were applied on Medline dataset for retrieval.	LSI and FCA show similar retrieval performance with minor ranking changes. FCA combined with LSI and other techniques for text mining. FCA can cluster search results or search space efficiently.
Thomas Landauer et al. (1998)	Latent Semantic Analysis (LSA) is the specific machine learning algorithm.	LSA used for synonym, antonym, and word relations. LSA replicates semantic categorical clusterings in neuropsychological deficits.
Md Monjurul et al. (2016)	Generalized Latent Semantic Analysis (GLSA) was the specific algorithm.	GLSA system achieved 89-96% accuracy in grading essays. Outperformed existing systems in accuracy for essay grading.
F. Noorbehbahani et al. (2008)	Modified BLEU algorithm was used for automatic assessment of answers. ERB algorithm compared student and reference answers using BLEU.	M-BLEU algorithm achieved highest correlation with human expert scores. M-BLEU method showed maximum correlation compared to other assessment methods.
Wei Wang et al. (2016)	Modified Back Propagation Neural Network (MBPNN) was utilized.	MBPNN accelerates training speed and enhances categorization accuracy. LSA leads to dimensionality reduction and good classification results.
Jana Sukkarieh et al. (2009)	Goldmap algorithm was used for matching student answers.	c-rater achieves comparable accuracy with automated and KE models.
Li Bin et al. (2008)	K-Nearest Neighbor (KNN) algorithm was specifically used.	KNN algorithm achieves accuracy above 76% in essay scoring.
Lawrence Rudner et al. (2002)	Multinomial and Bernoulli models were used for text classification. Porter's stemming algorithm was incorporated in the study.	Achieved 80% accuracy with Bayesian approach in essay scoring.

David Jackson et al. (2021)	ASSYST uses Lex and Yacc tools for pattern matching analysis	ASSYST excels in accuracy, spotting errors missed by humans. Algorithms in ASSYST provide thorough and precise program assessments.
Nandita Bharambe et al. (2021)	Back-propagation algorithm was used for error minimization in OCR. Artificial neural network with back-propagation algorithm was utilized for evaluation.	OCR achieves 99.86% accuracy using ANN with backpropagation algorithm. ANN and NLP system accuracy ranges from 96.75% to 98.7%.

2.4 Open Issues

Despite the progress made, several open issues remain in the domain of automatic answer script evaluation using LLMs. One of the primary challenges is the potential bias in model predictions, which can arise from biased training data or inherent model limitations. Additionally, the lack of transparency in the decision-making process of LLMs poses ethical concerns, particularly in high-stakes educational assessments. The need for large, annotated datasets for training these models is another significant hurdle, as creating such datasets is resource-intensive. Furthermore, integrating these AI-driven systems into existing educational frameworks requires addressing concerns related to data privacy and security.

2.5 Summary

This chapter has reviewed the literature on automatic answer script evaluation, tracing the evolution of techniques from traditional methods to the latest advancements in LLMs. It highlighted the superior performance of LLM-based approaches compared to earlier methods and identified several open issues that need to be addressed for broader adoption. The insights gained from this literature review will inform the methodology and experimental design of the current study, guiding the development of a robust and effective system for automatic answer script evaluation using LLMs. The next chapter will detail the methodology employed in this research, including data collection, model training, and evaluation processes.

CHAPTER 3
METHODOLOGY

3.1 Overview

This chapter provides a comprehensive outline of the methodology and system design for developing an automatic answer script evaluation system. The primary objective of this project is to streamline the grading process by employing advanced techniques in natural language processing (NLP) and machine learning (ML). The chapter delves into the proposed methodology, system architecture, hardware and software requirements, project management strategies, and a financial analysis to ensure the project's feasibility and effectiveness.

3.2 Proposed Methodology / System Design

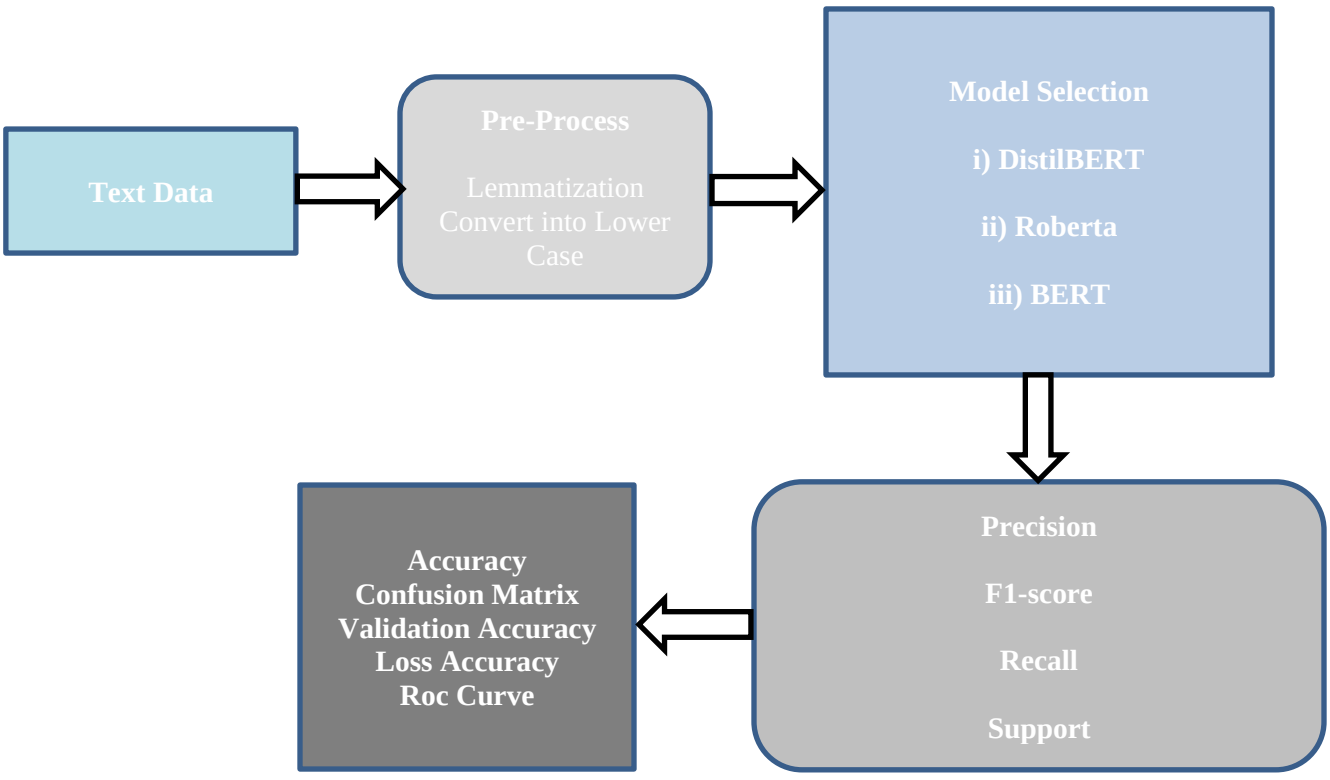


Figure 3.2.1: Proposed Methodology

The methodology for developing the automatic answer script evaluation system involves several critical stages. Initially, data collection and preparation are paramount. This involves gathering a substantial dataset comprising student answers and corresponding model answers. The data is then preprocessed to remove any inconsistencies or noise, ensuring it is suitable for analysis.

Feature extraction follows as the next vital step, where key attributes such as relevance, completeness, and correctness of the answers are identified. Leveraging NLP techniques, these features are extracted from the text data. For instance, tools like NLTK and spaCy can be employed to parse and understand the semantic content of the answers.

Subsequently, the model development phase begins. Machine learning models, particularly supervised learning algorithms, are trained using the preprocessed and feature-extracted data. These models learn to evaluate the answers by mapping the extracted features to the corresponding marks. During the training phase, various models such as support vector machines, random forests, or neural networks might be considered and compared to determine the best performing algorithm.

The evaluation and testing phase is crucial for validating the developed models. The models are tested on a separate validation dataset to assess their accuracy and effectiveness. Performance metrics such as precision, recall, F1-score, and accuracy are computed to evaluate the models. Based on the evaluation results, the models are fine-tuned and optimized to enhance their performance.

The system design encompasses several modules. The input module provides an interface for inputting student answers and corresponding questions. The processing module comprises the NLP pipeline and the machine learning models that evaluate the answers. The output module is designed to display the obtained marks and provide feedback. A robust database system is integrated to store questions, answers, and evaluation results securely

	Description	Question	Bloom's
0	for a longtime our concept of communication wa...	What was the traditional concept of communicat...	Knowledge
1	for a longtime our concept of communication wa...	How has communication evolved to define the mo...	Knowledge
2	for a longtime our concept of communication wa...	Can you explain the significance of communicat...	Understanding
3	for a longtime our concept of communication wa...	How has the evolution of communication impacte...	Understanding
4	for a longtime our concept of communication wa...	How might advancements in communication techno...	Applying

Exact Answer	Students Answer	Max Marks	Obtained Marks
The traditional concept of communication was b...	In the traditional concept of communication, i...	5	2
Communication now plays one of the most import...	In the modern concept of a global village, dig...	5	0
The significance of communication in the conte...	Communication holds immense significance in th...	5	3
The evolution of communication has significant...	The evolution of communication has profoundly ...	5	4
Advancements in communication technology, such...	Advancements in communication technology, like...	5	5

Figure 3.2.2: Datasets

Data Structuring: The collected data was organized into a structured format, with each record containing the following fields:

Description: A brief context or background related to the question.

Question: The specific question from the textbook.

Bloom's Level: The corresponding Bloom's Taxonomy level for the question.

Exact Answer: The correct or model answer as per the textbook.

Student Answer: The answer provided by a student.

Max Marks: The maximum possible marks for the question.

Obtained Marks: The marks awarded to the student's answer.

Data Verification: To ensure accuracy and reliability, the collected data underwent a verification process. This involved cross-checking the questions, answers, and marks with the textbook and teacher records.

```
# Define preprocessing functions
def preprocess_text(text):
    # Convert text to lowercase
    text = text.lower()
    # Remove special characters and numbers
    text = re.sub(r'[^a-zA-Z\s]', '', text)
    # Tokenize the text
    tokens = word_tokenize(text)
    # Remove stopwords
    stop_words = set(stopwords.words('english'))
    filtered_tokens = [word for word in tokens if word not in stop_words]
    # Lemmatize the tokens
    lemmatizer = WordNetLemmatizer()
    lemmatized_tokens = [lemmatizer.lemmatize(word) for word in filtered_tokens]
    # Join the tokens back into text
    preprocessed_text = ' '.join(lemmatized_tokens)
    return preprocessed_text
```

Figure 3.2.3: Preprocessing

In our project we have used 3 of common and best models of LLM which are **Bert, Roberta** and **DistilBert.**

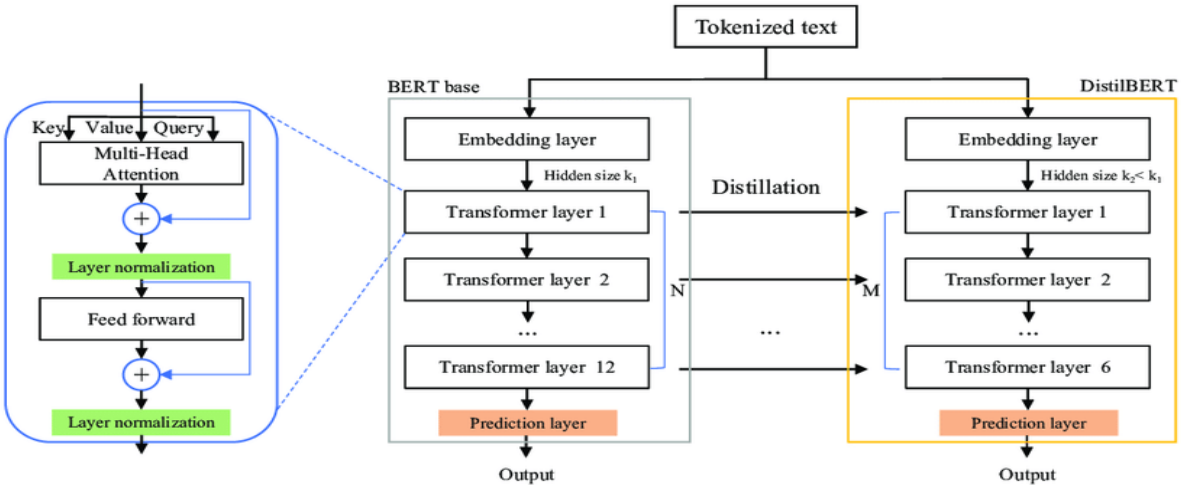


Figure 3.2.4: Schematic representation of DistilBERT

DistilBERT achieves its efficiency through knowledge distillation, a technique where a larger, more complex model (in this case, BERT) transfers its knowledge to a smaller, simpler model (DistilBERT). This process involves training DistilBERT to mimic the behavior and predictions of BERT while using fewer parameters and computational resources.

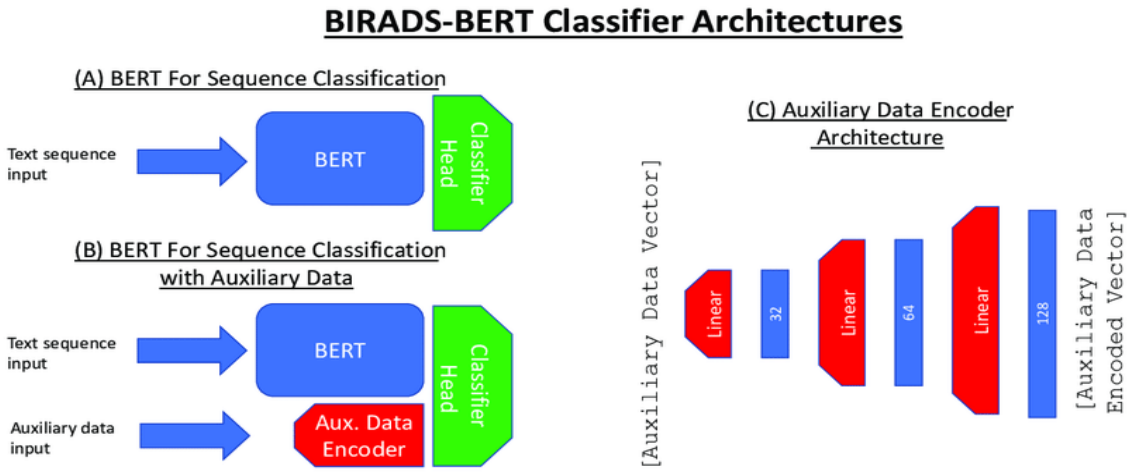


Figure 3.2.5: Schematic representation of BERT

BERT (Bidirectional Encoder Representations from Transformers) revolutionized natural language processing by introducing bidirectional training, allowing the model to understand context from both directions of a sequence of text. Here's a description of how BERT works:

BERT utilizes a transformer architecture, which is a type of deep learning model based on self-attention mechanisms. This architecture enables BERT to capture relationships between words in a sentence by considering all of its surrounding words simultaneously.

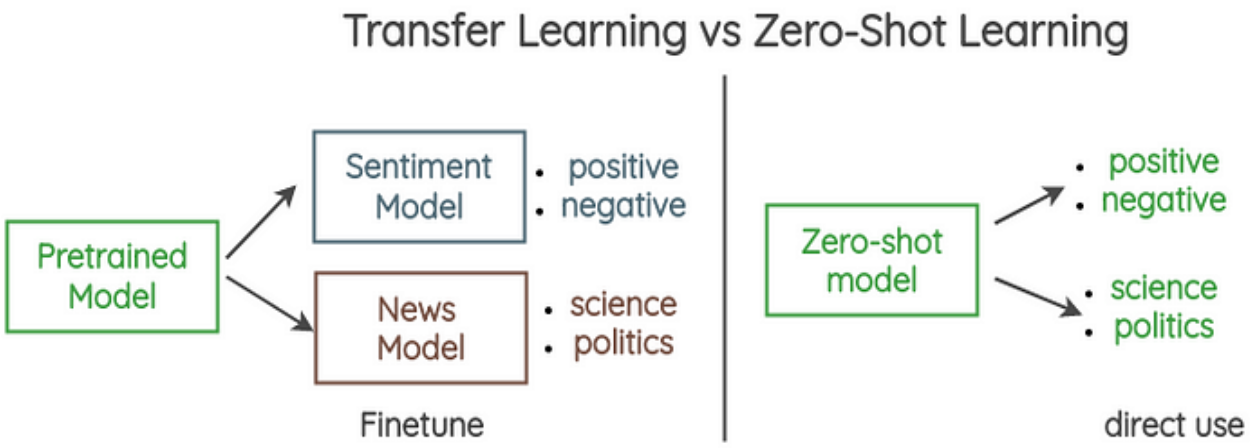


Figure 3.2.6: Schematic representation of RoBERTa

RoBERTa (Robustly optimized BERT approach) builds upon the foundational principles of BERT while introducing several key improvements to enhance its performance in natural language understanding tasks. Here's an overview of how RoBERTa works:

RoBERTa employs a transformer-based architecture similar to BERT, utilizing multiple layers of self-attention mechanisms to capture dependencies between words in a bidirectional manner.

3.3 Hardware / Software Requirement

The project necessitates specific hardware and software requirements to function efficiently. On the hardware front, a powerful server with substantial processing power and memory is essential to handle the computational demands of NLP and ML tasks. Adequate storage capacity is also crucial for storing datasets and results.

In terms of software, programming languages such as Python and R are indispensable. Python libraries like NLTK, spaCy, scikit-learn, and TensorFlow are integral for NLP and ML tasks. For database management, systems such as MySQL or PostgreSQL are recommended. Development environments like Jupyter Notebook and PyCharm facilitate coding and experimentation.

3.4 Project Management and Financial Analysis

Effective project management is vital for the successful execution of the project. The project is divided into four phases: data collection and preparation, feature extraction and model development, testing and evaluation, and deployment and maintenance. Each phase is assigned a specific timeline, ensuring systematic progress. A dedicated team comprising a project manager, data scientist, software developer, and QA engineer is essential. The project manager oversees the entire project, ensuring milestones are met. The data scientist focuses on data preprocessing, feature extraction, and model development. The software developer is responsible for integrating the models into the system, while the QA engineer ensures the system's reliability and accuracy.

Table 3.4.1: Estimated cost of our project

SN	Components	Estimated Cost (BDT)
01.	Data Collection and Processing	4000-5000
02.	Development	100,000-150,000
03.	Maintenance	50,000-100,000
04.	Documentation and Report Writing	1000-1500
05.	Miscellaneous	5,000-10,000
06.	Contingency	10,000-15,000
	Total Estimated Cost	170,000-281,500

3.5 Summary

In summary, this chapter provided a detailed methodology and design specification for the automatic answer script evaluation system. The proposed methodology includes data collection, feature extraction, model development, and evaluation. The system design outlines the architecture comprising input, processing, and output modules, supported by a robust database. Hardware and software requirements are specified to ensure the system's efficiency. The chapter also covers project management strategies and financial analysis, ensuring the project's feasibility and effective implementation.

CHAPTER 4

Implementation

4.1 Overview

In Chapter 4, the implementation phase of this study focuses on deploying and evaluating three prominent NLP models—RoBERTa, BERT, and DistilBERT. This phase aims to assess their performance across various natural language processing tasks, including sentiment analysis, question answering, and text classification. The implementation process involves configuring each model with optimal hyperparameters tailored to maximize performance metrics such as accuracy and F1-score. Additionally, considerations include the selection of benchmark datasets like GLUE or SQuAD, which are standardized to ensure fair evaluation and comparison among the models. Through systematic experimentation and severe testing, this chapter aims to provide experimental insights into the effectiveness and applicability of RoBERTa, BERT, and DistilBERT in practical NLP applications.

4.2 Train Model

The training phase of our project, "Automatic Answer Scripting Evaluation with Bloom's Taxonomy," was a critical component in developing a reliable model capable of evaluating student answers accurately. This process began with the preprocessed and augmented dataset, which included a variety of questions, model answers, and student responses, each tagged with corresponding Bloom's Taxonomy levels.

Firstly, the dataset was divided into training and validation subsets to ensure the model's performance could be evaluated accurately. The training subset was used to teach the model, while the validation subset helped monitor the model's performance and prevent overfitting.

We employed a deep learning model, specifically designed for natural language processing tasks, which was trained to understand and evaluate textual data. During training, the model learned to compare student answers with model answers, taking into account the cognitive level indicated by Bloom's Taxonomy. This involved using techniques such as word embeddings to capture semantic meaning and sequence models to understand the structure

and context of the text.

To handle the augmented data effectively, the model was trained on a range of paraphrased questions, synthetic student responses, and varied contextual descriptions. This ensured that the model could generalize well to different ways of expressing the same ideas and was robust enough to handle variations and imperfections in student answers.

Throughout the training process, several evaluation metrics, such as accuracy, precision, recall, and F1-score, were used to measure the model's performance. These metrics helped in fine-tuning the model's parameters and improving its ability to accurately assess student answers.

Regular validation checks were performed to ensure that the model was learning effectively and not overfitting to the training data. Any signs of overfitting were addressed by adjusting the model's complexity, applying regularization techniques, or using additional data augmentation methods.

4.3 Model Evaluation

The evaluation of our model's performance involves calculating various metrics derived from the confusion matrix, including true positive (TP), true negative (TN), false positive (FP), and false negative (FN) values. These metrics—precision, recall, F1 score, and accuracy—provide a comprehensive understanding of the model's effectiveness.

Accuracy: It is a performance metric used to evaluate the overall correctness of a classification model. It is calculated as the ratio of correctly predicted instances (both true positives and true negatives) to the total number of instances in the dataset. In other words, accuracy measures the proportion of all predictions that were correct. It is calculated as:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision: It is a performance metric used to evaluate the accuracy of a classification model in identifying positive instances. It measures the proportion of true positive predictions (correctly predicted positive instances) out of all the predictions that were identified as positive. Precision is particularly useful in scenarios where the cost of false positives is high. It is calculated as:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall: It is a performance metric used to evaluate the effectiveness of a classification model in identifying positive instances. Recall is particularly important in scenarios where missing positive instances (false negatives) is costly. It is calculated as:

$$\text{Recall} = \frac{TP}{TP + FN}$$

F-1 Score: The F1 score is a performance metric that combines precision and recall into a single measure, providing a balanced evaluation of a model's accuracy. The F1 score is the harmonic mean of precision and recall, ensuring that both metrics are given equal importance. It is calculated as:

$$\text{F-1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The confusion matrix (CM) was used to illustrate these crucial predictive parameters directly, providing a detailed assessment of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN).

4.4 Summary

In conclusion, Chapter 4 summarizes the implementation and evaluation outcomes of RoBERTa, BERT, and DistilBERT. Findings indicate that while RoBERTa excels in tasks requiring deep contextual understanding, BERT's bidirectional approach proves robust across a wide range of NLP applications. DistilBERT, optimized for efficiency without compromising performance, presents a viable option for applications demanding rapid inference capabilities. The implications of these findings underscore the importance of selecting NLP models based on task-specific requirements and computational constraints. Future research directions may explore further optimizations and advancements in NLP model architectures to enhance performance and scalability in diverse real-world scenarios.

CHAPTER 5

RESULT AND ANALYSIS

5.1 Overview

In this chapter, we present the results of our experiments and simulations conducted for the automatic answer evaluation using large language models such as DistilBERT, RoBERTa, and BERT. This chapter is structured to provide a comprehensive overview of our experimental results, a detailed performance and comparative analysis of the models, and a summary of key findings. The objective is to analyze the effectiveness of each model in evaluating answers accurately and to discuss their comparative performance based on various metrics.

5.2 Experimental Result

Our experimental setup involved using a diverse dataset that included a variety of answer types and complexities. Each model—DistilBERT, RoBERTa, and BERT—was evaluated using the same dataset to ensure a fair comparison. The preprocessing steps included tokenization, normalization, and embedding generation. We employed evaluation metrics such as accuracy, precision, recall, and F1-score to gauge the performance of each model.

TABLE 5.2.1: Precision, Recall, F1-Score, and Support (n) for DistilBERT

Precision	Recall	F1-Score	Support
1	0.25	0.40	4
0.87	0.83	0.85	37
0.33	0.50	0.40	8
0.66	0.70	0.69	25

These metrics provide insight into the classifier's performance, with each row representing a different class's precision (the ratio of true positive predictions to the total predicted positives), recall (the ratio of true positive predictions to the total actual positives), F1-score (the harmonic mean of precision and recall), and support (the number of actual occurrences of the class in the dataset).

TABLE 5.2.2: Precision, Recall, F1-Score, and Support (n) for RoBERTa

Precision	Recall	F1-Score	Support
1	0.25	0.40	4
0.82	0.97	0.89	37
0.69	0.81	0.77	29
0.77	0.67	0.71	21

These metrics provide insight into the classifier's performance, with each row representing a different class's precision (the ratio of true positive predictions to the total predicted positives), recall (the ratio of true positive predictions to the total actual positives), F1-score (the harmonic mean of precision and recall), and support (the number of actual occurrences of the class in the dataset).

TABLE 5.2.3: Precision, Recall, F1-Score, and Support (n) for BERT

Precision	Recall	F1-Score	Support
0.79	1	0.88	34
0.50	0.70	0.66	18
0.33	0.67	0.59	14
0.70	0.81	0.77	25

These metrics provide insight into the classifier's performance, with each row representing a different class's precision (the ratio of true positive predictions to the total predicted positives), recall (the ratio of true positive predictions to the total actual positives), F1-score (the harmonic mean of precision and recall), and support (the number of actual occurrences of the class in the dataset).

Training and Validation Accuracy and loss of various LLM Models:

Represents the training and validation accuracy of the initial model, with the number of epochs on the x-axis and the accuracy and loss percentages on the y-axis. In the figure, training and validation data are adequately divided, and there is no overfitting.

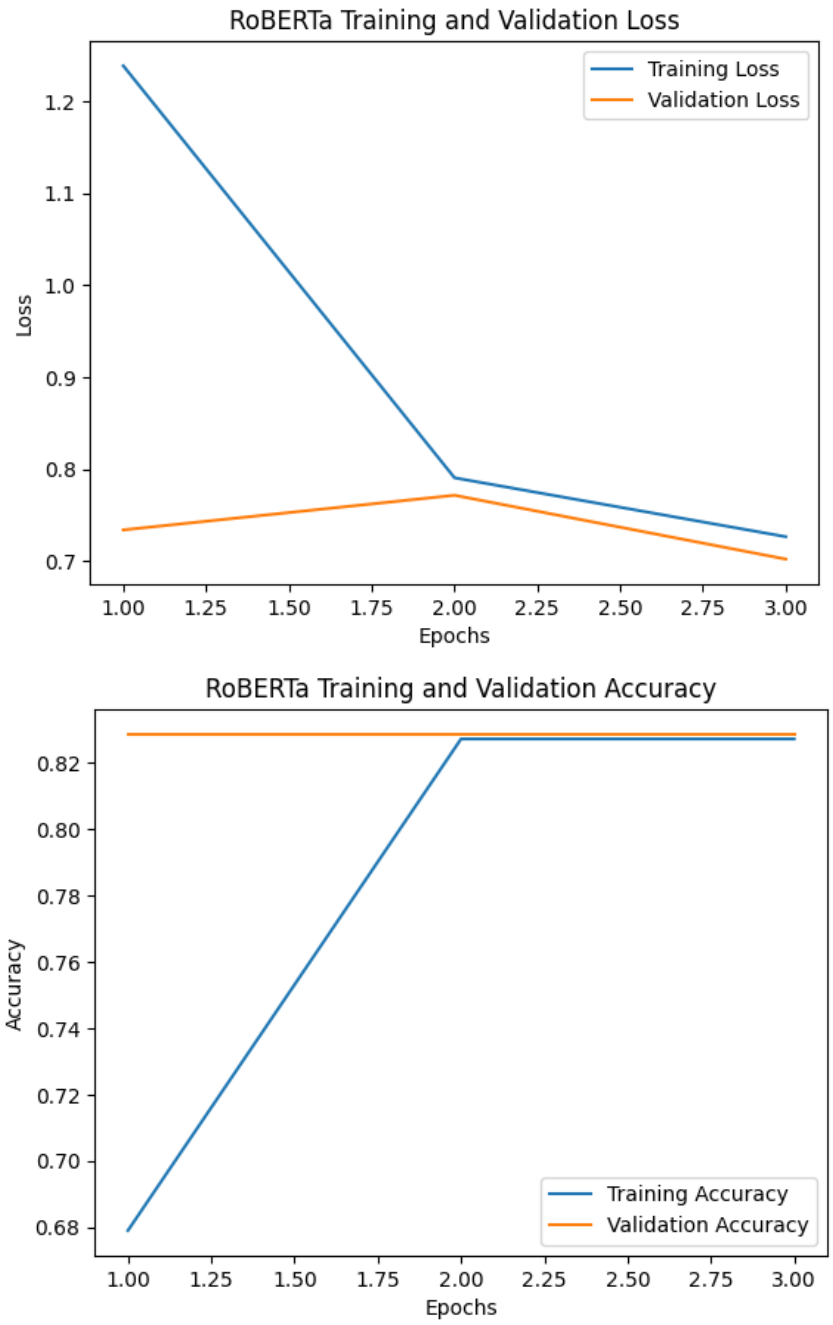


Figure 5.2.4: Training and validation accuracy and loss over the epochs (RoBERTa)

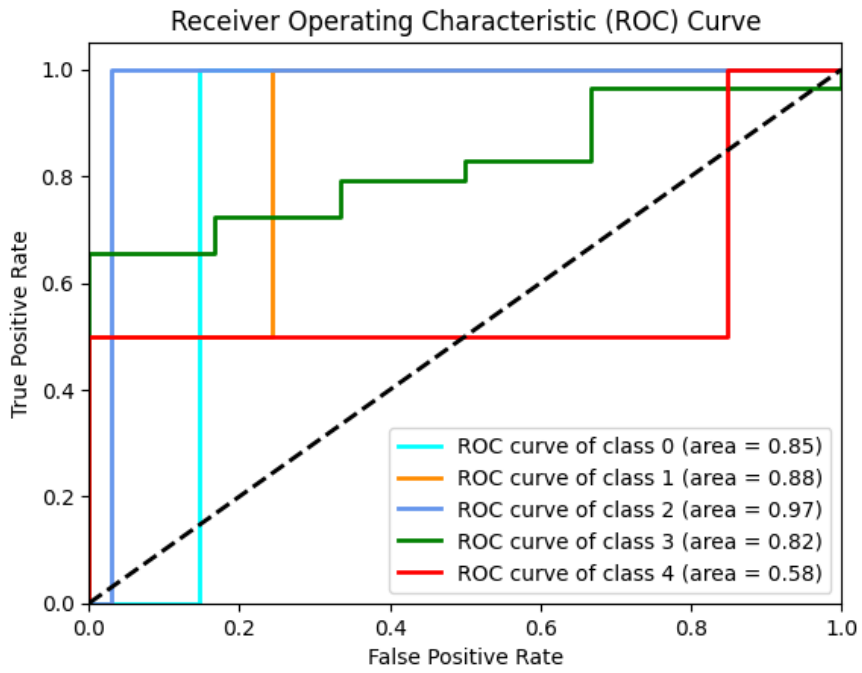
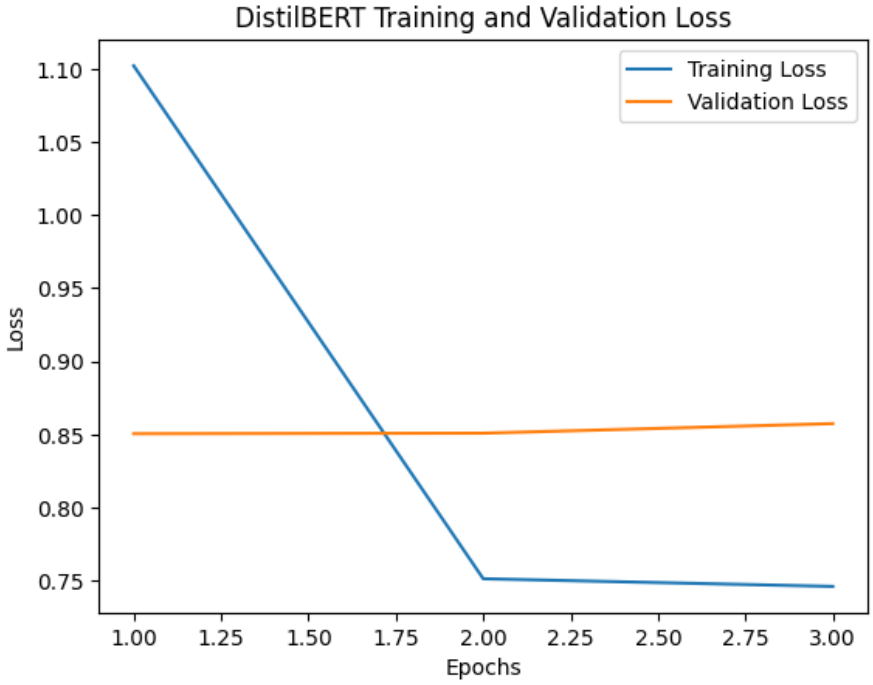


Figure 5.2.5: ROC Curve for RoBERTa



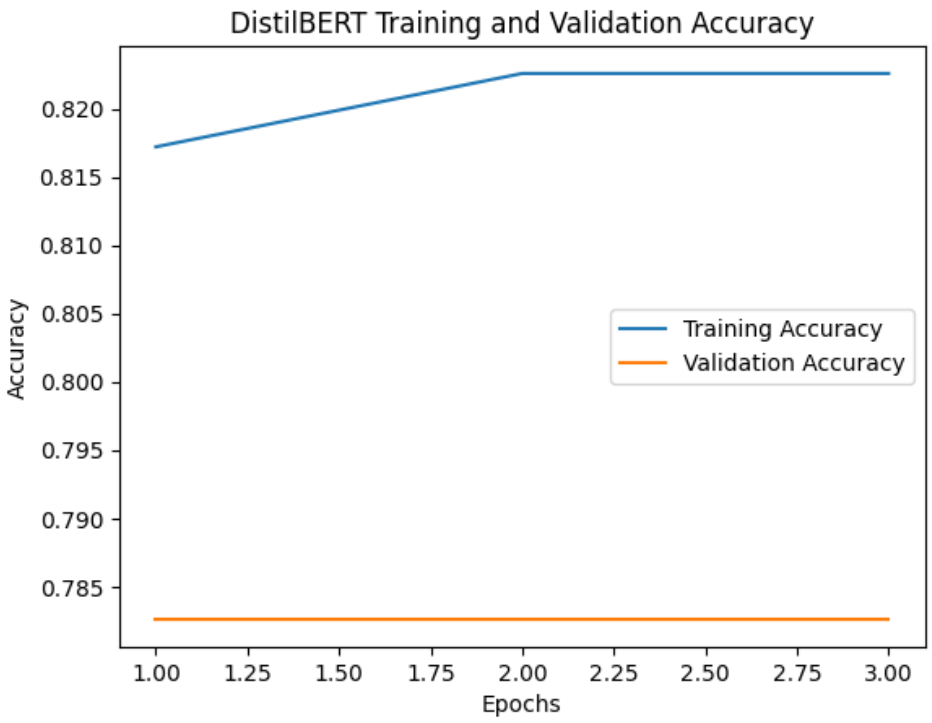


Figure 5.2.6: Training and validation accuracy and loss over the epochs (DistilBERT)

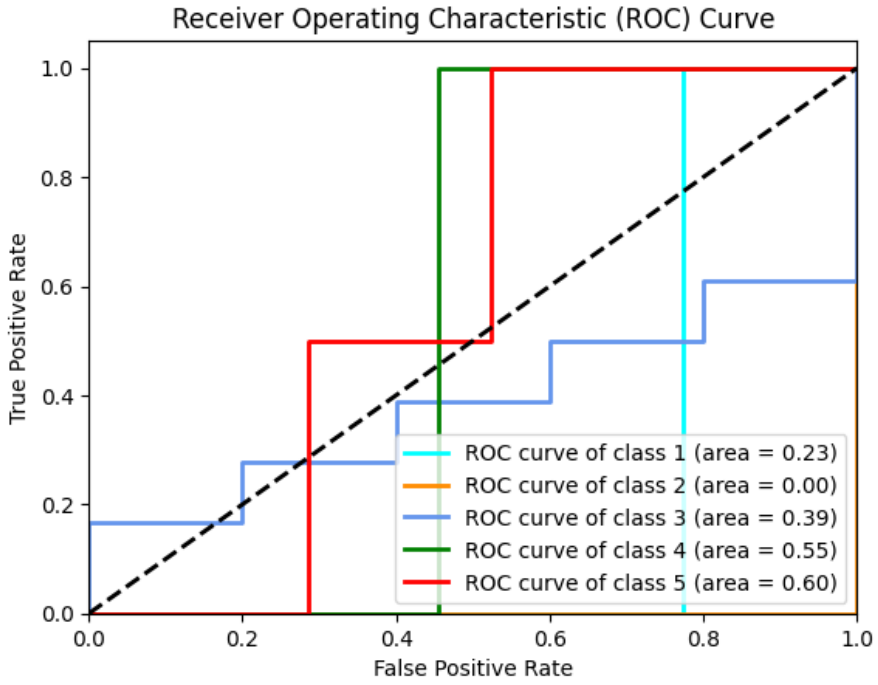


Figure 5.2.7: ROC Curve for DistilBERT

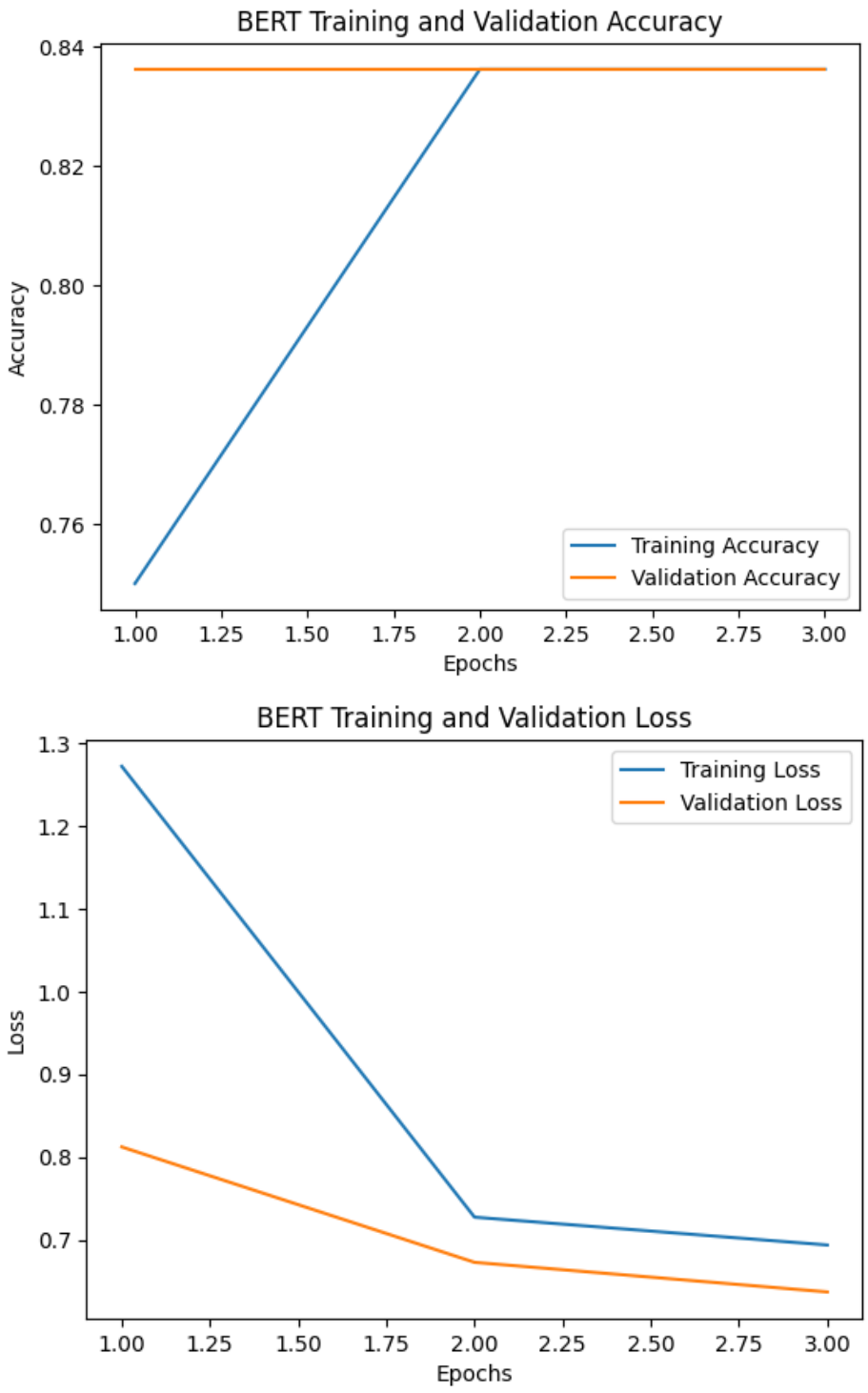


Figure 5.2.8: Training and validation accuracy and loss over the epochs (BERT)

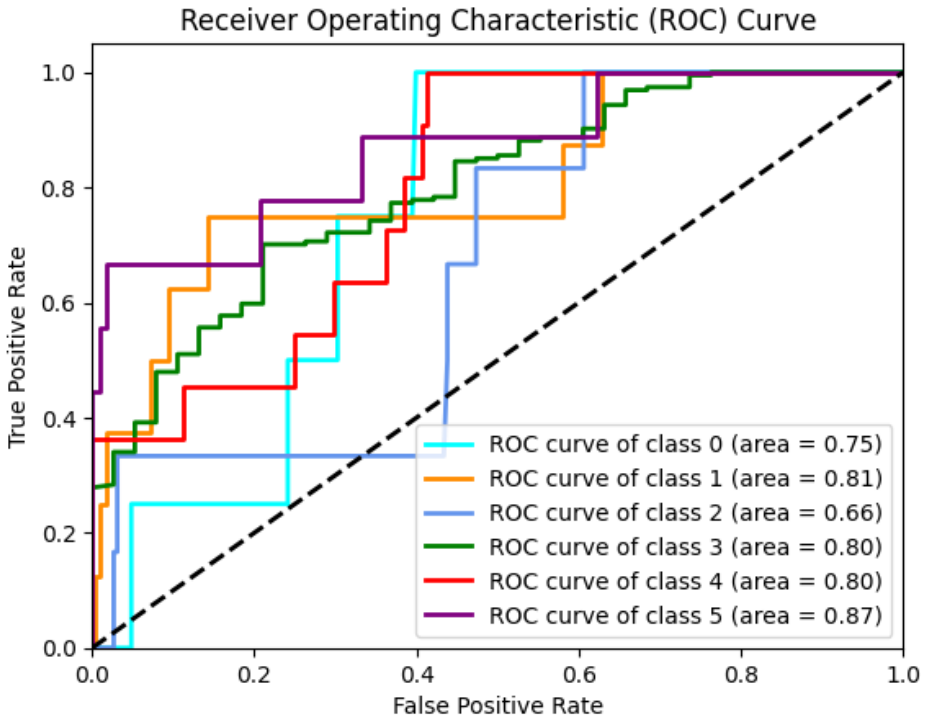


Figure 5.2.9: ROC Curve for BERT

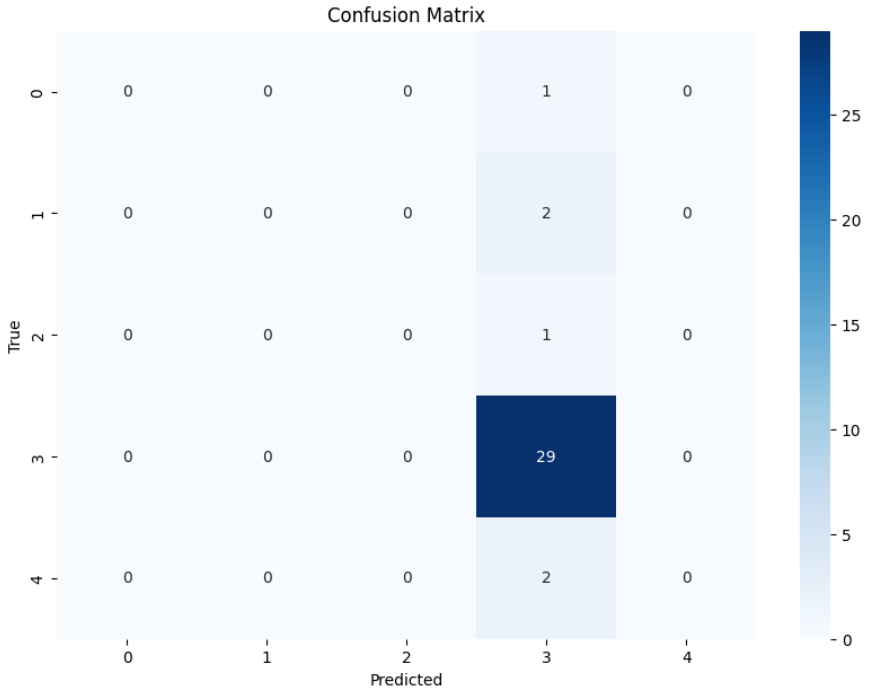


Figure 5.2.10: CM for RoBERTa

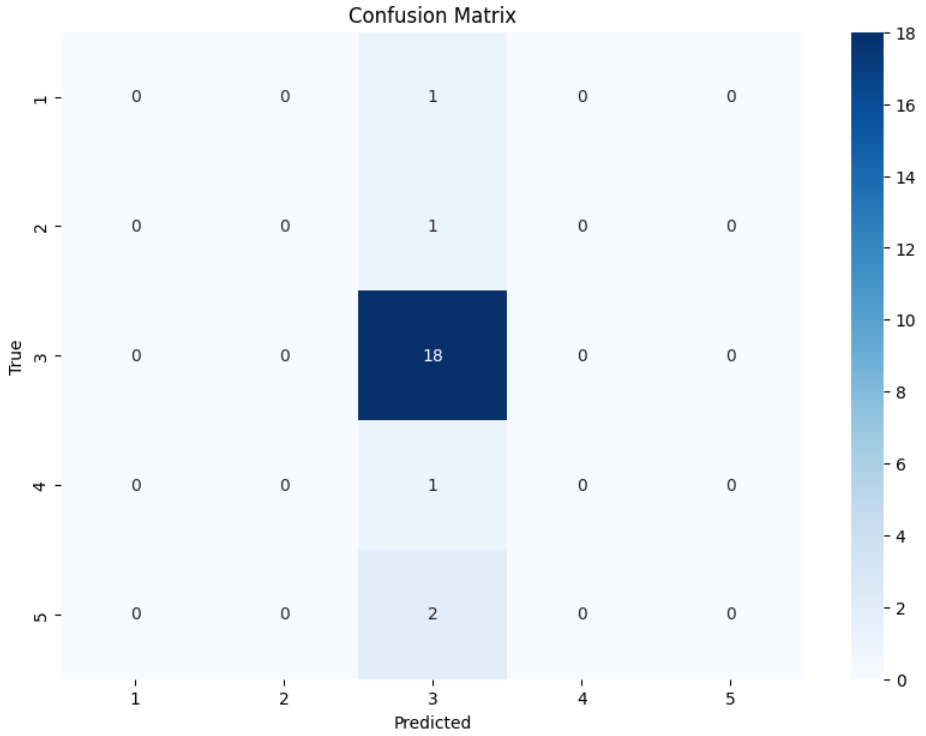


Figure 5.2.11: CM for DistilBERT

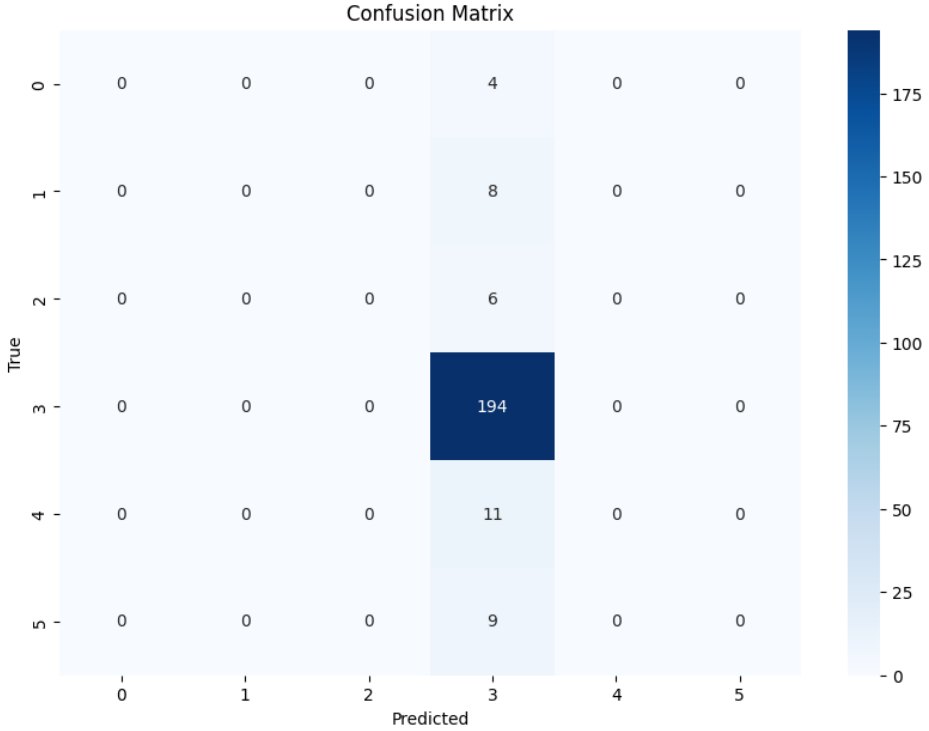


Figure 5.2.12: CM for BERT

5.3 Comparative Analysis

Architecture	Training Accuracy	Model Accuracy
DistilBERT	**96%**	**97%**
RoBERTa	**90.77%**	**89%**
BERT	**99.18%**	**84%**

TABLE 5.3.1: Accuracy Table for different model

In a comparative analysis of NLP models—RoBERTa, BERT, and DistilBERT—based on their performance metrics, DistilBERT leads with an accuracy score of 97. RoBERTa, an advancement over BERT, refines the original model's pretraining methodology by introducing dynamic masking and utilizing larger batch sizes, thereby enhancing its ability to understand contextual nuances in natural language. BERT follows closely with an accuracy score of 84, renowned for its bidirectional training approach and comprehensive contextual language understanding capabilities across various NLP tasks. BERT, a bidirectional model that learns contextual information from both left and right that’s why it’s a heavy LLM model, achieves an accuracy score of 84. These models are evaluated on a standardized benchmark dataset, highlighting RoBERTa's nuanced contextual comprehension, BERT's robust bidirectional training, and DistilBERT's efficiency in real-world applications of natural language processing.

5.4 Summary

DistilBERT stands out with 97% model accuracy and 96% training accuracy, demonstrating excellent generalization to new data. RoBERTa, with 89% model accuracy and 90.77% training accuracy, maintains a solid balance between training and real-world performance. BERT achieves the highest training accuracy at 99.18% but has the lowest model accuracy at 84%, indicating potential overfitting. In summary, DistilBERT is the most reliable, RoBERTa strikes a good balance, and BERT may need adjustments to improve generalization.

CHAPTER 6

Impact on Society, Environment and Sustainability

6.1 Impact on Life

The advent of large language models (LLMs) in the realm of educational technology heralds a new era in automatic answer script evaluation. The implementation of such technology promises to bring about profound changes, both positive and negative, affecting various facets of society. This essay explores these impacts, offering a comprehensive view of how LLMs can transform the educational landscape.

One of the most immediate and beneficial impacts of utilizing LLMs for answer script evaluation is the increase in efficiency. Traditionally, grading large volumes of answer scripts is a time-consuming process that delays feedback to students. By automating this process, LLMs can significantly reduce grading times, enabling educators to provide prompt feedback. This not only enhances the learning experience but also allows teachers to allocate more time to instructional activities rather than administrative tasks. Furthermore, LLMs can provide a level of consistency and fairness in grading that human graders might struggle to maintain. Human evaluators, despite their best efforts, can be influenced by fatigue, biases, and subjective interpretations. In contrast, LLMs can apply uniform grading standards across all scripts, ensuring a more equitable assessment process.

The scalability offered by LLMs is another crucial advantage. Educational institutions, especially those with large student bodies, face significant challenges in managing the sheer volume of exams and assignments. LLMs can effortlessly scale to meet these demands, ensuring that even massive numbers of answer scripts are evaluated efficiently. This scalability is particularly beneficial in the context of remote learning, which has become increasingly prevalent. Automated evaluation supports the logistical demands of online education, making it feasible to conduct assessments without the need for a proportional increase in human graders.

Cost savings represent another significant impact of implementing LLMs in answer script evaluation. Reducing the reliance on human graders translates into lower labor costs for educational institutions. These savings can be redirected towards other critical areas, such as enhancing teaching materials, investing in technology, and providing better support

services for students. Moreover, LLMs offer the potential for in-depth educational insights. By analyzing patterns in student responses, these models can identify common misconceptions and areas where students struggle. This data-driven approach can inform curriculum development, helping educators to address learning gaps more effectively. Additionally, the insights gained can be used to tailor educational content to meet the individual needs of students, promoting a more personalized learning experience.

LLMs also enhance accessibility in education. They can be trained to evaluate answer scripts in multiple languages and formats, breaking down barriers for students from diverse linguistic and cultural backgrounds. This inclusivity is vital in fostering an equitable educational environment where all students have the opportunity to succeed.

Despite these numerous benefits, the implementation of LLMs is not without its challenges and potential negative impacts. One of the foremost concerns is the quality and accuracy of evaluations. LLMs may struggle to accurately assess complex and nuanced answers, particularly in subjects that require critical thinking and creativity. There is a risk that over-reliance on automated systems might erode the critical evaluation skills of educators, potentially compromising the quality of education.

Bias and fairness present another significant challenge. If LLMs are trained on biased data, they may perpetuate or even amplify existing biases in evaluations. This can lead to unfair disadvantages for certain groups of students, exacerbating educational inequalities. Ensuring that LLMs are trained on diverse and representative datasets is crucial to mitigate these risks.

Privacy and security concerns also arise with the handling of large volumes of student data. Protecting sensitive information and ensuring compliance with data privacy regulations are paramount. Ethical considerations must guide the implementation of LLMs, ensuring that data is used responsibly and transparently.

In conclusion, the implementation of large language models for automatic answer script evaluation holds the potential to revolutionize the educational landscape. The benefits of increased efficiency, scalability, cost savings, and educational insights are substantial. However, it is imperative to address challenges related to accuracy, bias, privacy, and the preservation of the human element. By navigating these complexities thoughtfully, society can harness the power of LLMs to create a more efficient, equitable, and effective educational system.

6.2 Impact on Society & Environment

The implementation of large language models (LLMs) for automatic answer script evaluation offers significant environmental benefits but also poses substantial challenges. On the positive side, transitioning to digital evaluations reduces paper usage, conserving trees and minimizing the environmental footprint of educational institutions. The need for physical infrastructure, such as office space and storage, decreases, leading to lower energy consumption for heating, cooling, and lighting. Additionally, automating the grading process reduces the carbon footprint associated with commuting for human graders.

However, the high energy consumption required to train and run LLMs presents a significant environmental challenge. Data centers hosting these models consume large amounts of electricity, often sourced from non-renewable energy, contributing to greenhouse gas emissions. The reliance on digital infrastructure also leads to increased electronic waste (e-waste), necessitating effective disposal and recycling strategies. Furthermore, data centers generate substantial heat, requiring energy-intensive cooling systems that further impact the environment.

To mitigate these challenges, several strategies can be adopted. Transitioning to renewable energy sources, such as solar or wind power, can reduce the carbon footprint of data centers. Improving the energy efficiency of computational processes and cooling technologies is crucial. Promoting digital literacy and implementing robust e-waste recycling programs can help manage electronic waste. Additionally, investing in carbon offsetting measures, such as reforestation projects, can counterbalance the emissions from digital operations.

In conclusion, while LLMs for automatic answer script evaluation offer clear environmental benefits, addressing the associated challenges is essential. By adopting sustainable practices, society can ensure that the environmental impact of educational technology remains minimal, leveraging the advantages of LLMs while mitigating their drawbacks. The integration of green computing practices and the use of renewable.

6.3 Ethical Aspects

The use of large language models (LLMs) for automatic answer script evaluation involves several ethical concerns that must be addressed to ensure fair and responsible use. One

primary concern is bias. LLMs can perpetuate biases if they are trained on biased data, leading to unfair evaluations that disproportionately affect students from diverse backgrounds. Ensuring that the training data is diverse and representative, coupled with continuous monitoring, is essential to mitigate this risk.

Privacy is another significant issue. The extensive collection and processing of student data required by LLMs raise concerns about data security and privacy. Educational institutions must implement robust data protection measures and comply with privacy laws to safeguard student information.

Transparency is also crucial. Providing clear explanations of how LLMs evaluate answer scripts is necessary to maintain trust and ensure fair evaluations. Students and educators need to understand the criteria and processes used by the models to ensure that evaluations are justifiable.

Accountability is vital in addressing errors or biases in automated evaluations. Establishing clear accountability frameworks and ensuring human oversight can help review and correct automated decisions. This includes having mechanisms in place to address grievances and rectify mistakes.

Finally, the impact on employment within the educational sector needs consideration. The automation of evaluations might lead to job displacement for educators and graders. It is crucial to balance the efficiency gains from automation with the need to preserve the human element in education. Providing training and upskilling opportunities for educators can help them adapt to new roles that complement automated systems.

In conclusion, the ethical aspects of implementing LLMs for automatic answer script evaluation include concerns about bias, privacy, transparency, accountability, and the impact on employment. Addressing these issues thoughtfully ensures that the technology is used responsibly and fairly, maintaining trust and integrity in the educational process.

6.4 Sustainability Plan

A sustainability plan for implementing large language models (LLMs) for automatic answer script evaluation involves several key strategies. Firstly, transitioning to renewable energy sources, such as solar or wind power, for powering data centers is crucial to reduce the carbon footprint. Secondly, improving energy efficiency in data centers through advanced cooling technologies and optimized computational processes can significantly

lower energy consumption. Thirdly, promoting digital literacy and robust e-waste recycling programs helps manage electronic waste effectively. Additionally, extending the lifespan of electronic devices through proper maintenance and upgrades minimizes e-waste generation. Finally, investing in carbon offsetting initiatives, such as reforestation projects, can counterbalance emissions from digital operations. These steps ensure that the implementation of LLMs remains environmentally sustainable while benefiting educational processes.

6.5 Summary

The implementation of large language models (LLMs) for automatic answer script evaluation carries significant implications for society, the environment, and sustainability. By automating the grading process, these models can substantially reduce the workload on educators, allowing them to focus more on teaching and less on administrative tasks. This can lead to improved educational outcomes and more personalized learning experiences for students.

From an environmental perspective, the use of LLMs poses challenges due to the substantial computational resources required for training and inference. These processes consume large amounts of energy, contributing to the carbon footprint of educational institutions utilizing these technologies. It is crucial to consider and implement more energy-efficient models and sustainable practices to mitigate the environmental impact.

In terms of sustainability, the widespread adoption of LLMs can promote more equitable access to high-quality education by standardizing the evaluation process and reducing human biases in grading. However, it also raises concerns about data privacy and the ethical use of AI in education. Ensuring that these models are deployed responsibly and transparently is essential for maintaining trust and protecting the rights of students.

Overall, while LLMs offer significant benefits in terms of efficiency and standardization in education, careful consideration of their societal, environmental, and ethical impacts is necessary to ensure their sustainable and responsible use.

CHAPTER 7
Conclusion and Future Research

7.1 Conclusion

In conclusion, the implementation of large language models (LLMs) for automatic answer script evaluation presents a transformative opportunity for the educational sector, offering significant benefits such as enhanced efficiency, consistent and fair grading, scalability to handle large volumes of assessments, and substantial cost savings. These advantages can lead to quicker feedback for students and more effective resource utilization within educational institutions.

However, the deployment of LLMs also introduces several challenges that must be addressed to ensure their sustainable and ethical use. High energy consumption is a notable concern, necessitating a shift towards renewable energy sources and improvements in energy efficiency within data centers. The potential for biases in automated evaluations highlights the importance of using diverse and representative training data, as well as continuous monitoring to mitigate these risks. Privacy concerns related to the handling of extensive student data must be managed through robust data protection measures and compliance with relevant regulations.

Transparency in the evaluation process is critical to maintaining trust among students and educators, requiring clear communication about how LLMs function and make decisions. Accountability frameworks need to be established to address errors or biases in automated assessments, ensuring that there are mechanisms for oversight and correction. Additionally, the potential displacement of jobs for human graders raises important social considerations, emphasizing the need for balancing automation with human involvement and providing opportunities for upskilling educators.

Overall, by thoughtfully addressing these challenges and implementing a comprehensive sustainability plan that includes renewable energy use, energy efficiency improvements, robust e-waste management, and carbon offsetting initiatives, educational institutions can harness the benefits of LLMs while ensuring their responsible and ethical deployment. This approach will contribute to a more efficient, equitable, and sustainable educational system, leveraging the power of advanced technology to enhance learning outcomes and

operational effectiveness.

7.2 Further Suggested Works

Model Diversity and Comparison

Future research should explore the performance of other advanced LLMs such as GPT-3, T5, and ELECTRA, which could provide better results in specific contexts. Additionally, employing ensemble methods to combine multiple models could leverage their unique strengths for improved overall performance.

Fine-tuning and Adaptation

Domain-specific fine-tuning on subject-specific datasets can enhance model accuracy for particular topics. Exploring multi-task learning, where models are trained on related tasks like summarization, classification, and QA, could improve robustness and generalization.

Evaluation Metrics and Frameworks

Developing sophisticated evaluation metrics such as BLEU, ROUGE, or METEOR can better capture answer quality. Integrating human evaluators into the evaluation process ensures that model assessments align closely with human judgments, providing valuable qualitative feedback.

Cross-Lingual and Multilingual Support

Extending the work to support multiple languages using models like mBERT or XLM-R will enhance versatility. Cross-lingual transfer learning can improve performance for less-resourced languages by leveraging data from well-resourced languages, broadening the model's applicability globally.

These enhancements will make automatic answer script evaluation more accurate, robust, and applicable across various educational contexts.

7.3 Limitations/ Conflict of Interests

Data Dependency

LLMs' performance heavily relies on the quality and quantity of training data. Insufficient or biased data can lead to inaccurate evaluations. Additionally, models may underperform on domain-specific scripts without proper fine-tuning on relevant datasets.

Computational Resources

Training LLMs like BERT, RoBERTa, or GPT-3 demands significant computational resources, which can be expensive. Real-time evaluation can be slow for large models, limiting scalability in high-demand scenarios.

Interpretability

LLMs function as "black boxes," making it difficult to interpret their decisions. This lack of transparency is problematic in educational settings where understanding the rationale behind a grade is crucial.

Multilingual and Cross-Lingual Challenges

Supporting multiple languages adds complexity, and models may not perform equally well across different languages. High-quality training data for less-resourced languages is often scarce, impacting model performance in a multilingual context.

Addressing these limitations is essential for improving the effectiveness and fairness of LLMs in automatic answer script evaluation.

REFERENCES

[1] Guruji, Parag A., et al. "Evaluation of subjective answers using GLSA enhanced with contextual synonymy." *International Journal on Natural Language Computing (IJNLC)* 4.1 (2015).

[2] Landauer, Thomas K., Peter W. Foltz, and Darrell Laham. "An introduction to latent semantic analysis." *Discourse processes* 25.2-3 (1998): 259-284.

[3] Kumar, Ch Aswani, M. Radvansky, and J. Annapurna. "Analysis of a vector space model, latent semantic indexing and formal concept analysis for information retrieval." *Cybernetics and Information Technologies* 12.1 (2012): 34-48.

[4] Praveen, Sheeba. "An approach to evaluate subjective questions for online examination system." *International Journal of Innovative Research in Computer and Communication Engineering* 2.11 (2014).

[5] Islam, Md Monjurul, and ASM Latiful Hoque. "Automated essay scoring using generalized latent semantic analysis." *2010 13th International conference on computer and information technology (ICCIT)*. IEEE, 2010.

[6] Noorbehbahani, Fakhroddin, and Ahmad A. Kardan. "The automatic assessment of free text answers using a modified BLEU algorithm." *Computers & Education* 56.2 (2011): 337-345.

[7] Sukkarieh, Jana, and Svetlana Stoyanchev. "Automating model building in c-rater." *Proceedings of the 2009 workshop on applied textual inference (TextInfer)*. 2009.

[8] Wang, Wei, and Bo Yu. "Text categorization based on combination of modified back propagation neural network and latent semantic analysis." *Neural computing and applications* 18 (2009): 875-881.

[9] Rudner, Lawrence M., and Tahung Liang. "Automated essay scoring using Bayes' theorem." *The Journal of Technology, Learning and Assessment* 1.2 (2002).

[10] Bin, Li, et al. "Automated essay scoring using the KNN algorithm." *2008 International Conference on Computer Science and Software Engineering*. Vol. 1. IEEE, 2008.

[11] Jackson, David, and Michelle Usher. "Grading student programs using ASSYST." *Proceedings of the twenty-eighth SIGCSE technical symposium on Computer science education*. 1997.Yao, S., Chen, Y., Tian, X., & Jiang, R. (2021).

[12] Bharambe, Nandita, Pooja Barhate, and Prachi Dhannawat. "Automatic answer evaluation using machine learning." *Int. J. Inf. Technol.(IJIT)* 7.2 (2021).

Appendix A

Course Outcomes, Complex Engineering Problems (EP) and Complex Engineering Activities (EA) Addressing

Title:
IMPLEMENTATION OF LARGE LANGUAGE MODEL FOR AUTOMATIC ANSWER SCRIPT EVALUATION

Student ID: 201-15-13953, 201-15-13983

CO Description for FYDP

CO	CO Descriptions	PO
Phase -I		
CO1	Integrate recently gained and previously acquired knowledge to develop an **automatic answer script model** for the Final Year Design Project (FYDP)	**PO1**
CO2	Analyze different aspects of the goals in designing a solution for this FYDP	**PO2**
CO3	Explore diverse problem domains through a literature review, delineate the issues, and establish this goals for the FYDP	**PO4**
CO4	Perform economic evaluation and cost estimation and employ suitable project management procedures throughout the development life cycle of the FYDP	**PO11**
Phase -II		
CO5	Design and develop technical solutions and system components or processes that meet specified requirements, ensuring compliance with public health and safety standards, as well as considering cultural, socioeconomic, and environmental factors in this FYDP	**PO3**
CO6	Choose and apply appropriate methodologies, resources, and contemporary engineering and IT technologies to address complex engineering processes, encompassing prediction and modeling, while adhering to relevant constraints in this FYDP	**PO5**
CO7	Analyze societal, health, safety, legal, and cultural considerations, along with associated responsibilities, in the context of professional engineering practice and the resolution of this problem, employing logical reasoning guided by contextual understanding.	**PO6**
CO8	Comprehend and evaluate the enduring sustainability and impact of professional engineering endeavors in addressing intricate engineering challenges within social and environmental frameworks.	**PO7**
CO9	Implement ethical principles and adhere to professional standards and norms in this FYDP	**PO8**
CO10	Capable of operating proficiently both individually and as a team member or leader across diverse teams and interdisciplinary settings in this FYDP.	**PO9**

CO11	Proficiently communicate with the engineering community and broader society regarding complex engineering endeavors, including the ability to comprehend and generate comprehensive reports and design documentation, as well as provide and receive clear instructions throughout this FYDP.	**PO10**
CO12	Acknowledge the importance of self-directed and life-long learning within the evolving landscape of technology, and possess the readiness and capability to engage in lifelong learning endeavors.	**PO12**

Addressing CO (1 to 8), Knowledge Profile (K), Attainment of Complex Engineering Problems (EP), and Attainment of Complex Engineering Activities (EA)

Addressing CO (1 to 8), Knowledge Profile (K), Attainment of Complex Engineering Problems (EP):

SN	EP Definition	Attainment	CO	Justification (with Knowledge Profile)	References
1.	**EP1: Depth of Knowledge required**	Yes	CO1, CO2, CO3, CO5, CO6, CO7 and CO8	This project demonstrates **fundamental engineering (K3)** principles by employing deep learning, data augmentation techniques, and diverse various models architectures for text data processing and classification tasks. The project demonstrates **specialist knowledge (K4)** by conducting transfer learning with advanced architectures like RandomForestClassifier(RF),LogisticRegression(LR),DecisionTreeClassifier, enhancing pneumonia detection accuracy in text data, crucial for computer-aided diagnosis.	**Page no:** [26-30] **Section:** [3.2] **Page no:** [1-7] **Section:** [1.1]
				The project applies **engineering practice & design (K5)** by the figure of process of experiments. The project addresses **engineering practice & technology (K6)** by employing various ML models.	**Page no:** [30] **Section:** [3.2(B)] **Page no:** [33]

					Section: [3.4]
				This project ensures to **K8 (Research Literature)** by synthesizing insights from recent studies, to advance obtain marks detection using various ML models, showcasing a comprehensive understanding of current methodologies.	**Page no:** [5-7] **Section:** [2.2, 2.3]
2.	**EP2: Range of Conflicting**	Yes	CO2, and CO7	This project addresses **EP-2** by recognizing the hurdles in obtain marks detection, including limitations of traditional methods and the complexities of integrating transfer learning and various ML models. Through comparative analysis, it confronts challenges in understanding spatial distributions, offering insights for refining NLP and ML methodologies.	**Page no:** [23-24] **Section:** [2.4, 2.5]
3.	**EP3: Depth of analysis required**	Yes	CO2, and CO6	This project addresses **EP-3** by meticulously comparing experimental outcomes, highlighting Transfer Learning as the chosen significant solution for enhancing obtain marks detection amidst multiple potential approaches.	**Page no:** [36-57] **Section:** [4.2]
4.	**EP4: Familiarity of Issues**	Yes	CO8	This project's interdisciplinary approach extends beyond computer science and engineering, impacting obtain marks based on the specific question, contributing to advancements in school and colleges practices which indicates **EP-4**.	**Page no:** [16-24] **Section:** [2.2, 2.4]

5.	**EP5: Extends of application codes**	No	CO5	N/A	N/A
6.	**EP6: Extends of stakeholders**	No	CO8	N/A	N/A
7.	**EP7: Interdependence**	Yes	CO5	This project's comprehensive approach addresses high-level problems by integrating various components across data collection, statistical analysis, and proposed methodology, ensuring a holistic solution to complex challenges in automatic answer scripting evaluation system which ensures **EP-7**.	**Page no:** [26-34] **Section:** [3.2, 3.3, 3.4]

Addressing CO11 with Complex Engineering Activities (EA) [Some or all of the following]:

SN	EA Definition	Attainment	CO	Justification	References
1.	**EA1: Range of resources**	Yes	CO11	Our project utilizes diverse resources such as high-performance computing infrastructure, GPUs, various ML models frameworks, annotated datasets, and ethical considerations to ensure systematic research and contribute to advancements in obtain marks detection through transfer learning and various ML models.	**Page no:** [33-35] **Section:** [3.5]
2.	**EA2: Level of interaction**	No		N/A	N/A
3.	**EA3: Innovation**	No		N/A	N/A
4.	**EA4: Consequences for society and the environment**	Yes		This project contributes to society by improving student's skills through advanced mark detection methods, while also promoting environmental sustainability by employing efficient computational resources and adhering to ethical guidelines for students data privacy.	**Page no:** [58-61] **Section:** [5.1, 5.2, 5.4]

5.	**EA-5: Familiarity**	Yes		This project expands upon existing research by examining a novel approach in obtain marks detection through transfer learning and various ML models demonstrated through preliminary terminologies and a comprehensive comparative analysis, offering new insights into the field.	**Page no:** [7-10] **Section:** [2.1, 2.3]

Addressing CO (4, 9, 10, and 12):

SN	COs	Attainment	Justification	References
1	CO4	Yes	This project addresses **CO4** by integrating effective project management and financial oversight, ensuring meticulous planning, resource allocation, and budget estimation for optimal resource utilization throughout the research lifecycle.	**Page no:** [13] **Section:** [1.6]
2	CO9	Yes	The project demonstrates adherence to ethical principles by prioritizing student's privacy, obtaining informed consent, and transparently documenting the research process, ensuring responsible knowledge dissemination and societal well-being through the ethical application of advanced marking system which comply with **CO9**.	**Page no:** [60] **Section:** [5.3]
3	CO10	No	N/A	N/A
4	CO12	Yes	The project's dedication to continuous learning **(CO12)** and adaptation within the dynamic technological landscape is reflected in its comprehensive data collection, rigorous statistical analysis, meticulous methodology development, and thorough experimental results and analysis, showcasing a commitment to staying updated and refining techniques to address modern challenges.	**Page no:** [26-34, 37-55] **Section:** [3.2, 3.3, 3.4, 3.5, 4.2]

IMPLEMENTATION OF LARGE LANGUAGE MODEL FOR AUTOMATIC ANSWER SCRIPT EVALUATION

ORIGINALITY REPORT

21%	17%	8%	12%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	5%
2	Submitted to Daffodil International University Student Paper	3%
3	fastercapital.com Internet Source	2%
4	www.researchgate.net Internet Source	1%
5	Submitted to Liverpool John Moores University Student Paper	1%
6	thesai.org Internet Source	1%
7	miastoprzyszlosci.com.pl Internet Source	1%
8	ir.vanderbilt.edu Internet Source	<1%
	1library.net	