DISTINGUISHING BETWEEN AI-GENERATED AND HUMAN-WRITTEN CONTENT IN THE MODERN DIGITAL LANDSCAPE USING DEEP LEARNING

By
MD. ATIQ MORSHED EMU
ID: 201-15-14262

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

Ms. Dristi Saha

Lecturer
Department of Computer Science and Engineering
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY
Dhaka, Bangladesh

January 13, 2025

APPROVAL

This Project titled **"DISTINGUISHING BETWEEN AI-GENERATED AND HUMAN-WRITTEN CONTENT IN THE MODERN DIGITAL LANDSCAPE USING DEEP LEARNING"**, submitted by Md. Atiq Morshed Emu, ID No: **201-15-14262** 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 12/13 January, 2025.

BOARD OF EXAMINERS

Dr. S.M Aminul Haque (SMAH) — **Chairman**
Professor and Associate Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Md. Abbas Ali Khan (AAK) — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Mr. Md. Aynul Hasan Nahid (AHN) — **Internal Examiner**
Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Dr. Md. Zulfiker Mahmud — **External Examiner**
Professor
Department of Computer Science and Engineering
Jagannath University

©Daffodil International University

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Ms. Dristi Saha, Lecturer**, Department of Computer Science and Engineering, Daffodil International University. We also declare that neither this projectnor any part of this project has been submitted elsewhere for the award of any degree or diploma.

Supervised by:

Dristi

Ms. Dristi Saha
Lecturer
Department of Computer Science and EngineeringDaffodil International University

Submitted by:

Md. Atiq Morshed Emu

Md. Atiq Morshed Emu
Student ID:201-15-14262
Department of Computer Science and EngineeringDaffodil International University

©Daffodil International University

ACKNOWLEDGEMENTS

This work would not have been possible without the support and contributions of many individuals over the past two semesters. We are deeply grateful to everyone who has assisted us in one way or another.

First, we express our heartfelt thanks and gratefulness to the almighty for His divine blessing making it possible for us to complete the **Final Year Design Project (FYDP)** successfully.

We are grateful and wish our profound indebtedness to **Ms. Dristi Saha, Lecturer**, Department of Computer Science and Engineering, Daffodil International University, Dhaka, Bangladesh. Deep knowledge and keen interest of our supervisor in the field of **Machine Learning, Image Processing, Mobile Application Development** to carry out this project. Her 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 heartfelt gratitude to the Head of the Department of Computer Science and Engineering, for his kind help in finishing our project and also to other faculty members and the staff of the Department of Computer Science and Engineering, Daffodil International University.

We would like to thank our entire course-mates at Daffodil International University, who took part in this discussion while completing the coursework.

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

ABSTRACT

This study focuses on the growing difficulty of recognizing text that produced by machines in a time when artificial intelligence is extensively used. The study proposes a novel method based on a Long Short-Term Memory (LSTM), GRU and Hybrid architecture to distinguish AI-generated content and human-written text with remarkable accuracy. By employing advanced techniques for text preprocessing, vectorization, and embedding, we achieved an efficient design with average computational demands. We tested the models on a large dataset, the model demonstrated outstanding performance. The best model achieved an accuracy of 98.31% and the best F1-score is 0.98. These findings show the outstanding ability of the model to generalize well on unseen data, proving the potential of using it in real-world applications. The model's stability and reliability are backed by highly similar outcomes of the training and validation phases with minimal overfitting due to excellent regularization strategies. The confusion matrices and the full classification reports gave in-depth insights into the model's strengths and weaknesses, thus enhancing its applicability.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter gives a summary of the subject of the research which includes an overview, motivation and objective, methodology, project outcome, report organization.

1.1 Introduction

Content creation has been enhanced by artificial intelligence (AI) in a number of domains in recent years. That means GPT-3 and GPT-4, and other AI-powered text generation technologies, can produce language that sounds human, and is often undistinguishable from human written content [1]. This development is advantageous and disadvantageous. With AI, you can use innovative content creation, document drafting, and automated repetitive writing tasks [2]. However, these questions revolve around content validity, disinformation propagation and intellectual property infringement [3]. And it's more and more important to be able to differentiate machine created content from human created content. There is more need than ever to check the validity and legitimacy of information in light of the flood of fake news and false information. Typically, conventional content verification techniques miss identifying the subtle differences between texts created by Artificial intelligence versus those produced by humans. Consequently, developing sophisticated natural language processing (NLP) methods to truly pinpoint the source of the content is critical [5]. To address this difficulty, this research surveys many advanced NLP approaches. We analyze the efficacy of machine learning algorithms like Random Forest, Logistic Regression, Decision Trees, Naive Bayes and some deep learning models such as Recurrent Neural Networks (RNN) and Convolution Neural Networks (CNN) [6]. Through this comparison, we attempt to understand the most suitable techniques for distinguishing between data generated by AI and human generated content. Our study is intended to supply a complete account of the methods and experiments used and results.

Then we take a look at the situation of AI text making today and the challenges it presents in the background part.

In the study methodology chapter, we describe how we create and evaluate various NLP models. The wider ramifications of our findings are addressed in the influence on society chapter, and the experimental implementation details the procedures we used to train and assess these models [7]. We aim to improve accuracy and reliability of content verification process towards more secure and trusted digital communications. In order to fight with false information and conserve information integrity in the digital world, we want to provide useful tools with advanced NLP techniques. [8]. The results of this study could have implications for many different stakeholders, including educators, journalists, policymakers and technology developers.

1.2 Motivation

In the last few years, AI has moved forward quickly, changing how the world produces, distributes and utilizes information. Different tools can generate text that seems and feels human written. While this technology has several advantages for business, education and production of content, there are disadvantages as well. Dissemination of misinformation and plagiarism, and crumbling of the faith in written communication, these are the concerns.

Now, teachers have a harder time deciding whether a piece of work was done by AI or by a student themselves. In sectors that need unique ideas, artificial intelligence programs are able to replicate writing styles and produce truly real material, thereby putting intellectual property at risk. With the deluge of AI generated information on the internet it becomes harder for readers to distinguish between what is reliable or authentic.

This study aims to resolve these very issues. It wants to create tools to differentiate AI generated writing from human authored. It does this by creating responsible content production, protecting academic integrity and a ethical application of AI.

1.3 Objectives

This study is justified by the explosive emergence of AI text technologies and

their distressing effect on society. While GPT-3 and GPT-4 represent an awesome step toward new possibilities that AI can create, they bring with them a whole new set of challenges to address—primarily being the transparency and accountability of online communications. Most existing AI text detection systems are not versatile and precise enough to generalize their results for datasets or domains. Current approaches are still only applicable to certain text types or contexts and don't adapt well to 'better' AI generated models of today.

And there are serious societal implications to unidentified AI generated text. AI tools can be misused by students in academic settings for the purpose of plagiarism, rendering work false. On the web, this AI driven content could be used to spread misinformation, which has the potential to harm on a global level. But industries based on original content may have trouble detecting copyright infringements or measure originality.

In this study, we fill these gaps by developing a robust methodology that can be used to analyze different types of textual data at high accuracy. This not only aims to enhance the cutting edge in AI-text detection, but also to equip educators, businesses and policy makers with tools that can be used safely and ethically in using AI technologies. This work helps close the gap between innovation and accountability, where the advancement of AI can be made to have a positive contribution to society without compromising authenticity, and trust.

1.4 Methodology

The research goals for the study are to develop an ML algorithm that can classify text as either written by AI or by human authors, based on cutting-edge NLP methodologies and computational methods. Currently, data is processed, models are built and evaluated on cloud platforms such as Kaggle, Google Colab as well as through libraries of Scikit-learn, TensorFlow, NLTK.

A text dataset of 487,236 samples from Hugging face and Kaggle are analyzed from the year 2024. LSTM model is selected because it is a model that can operate on sequential data, metrics including accuracy, precision, recall, F1-score.

The method entails data preprocessing (tokenization and removal of stop words)

data partitioning into training 70%, validation 15%, and finally the test set 15%, LSTM, GRU, and the hybrid model. The proposed model is developed with the Adam optimizer and the binary cross-entropy loss function with the use of early stopping to prevent overlearning. In the case of the final models, the accuracy is evaluated utilizing confusion matrix.

Requirements: It has a configuration of Intel i5 processor, 16GB RAM and GTX1650ti GPU. Techniques employed are Python features (NumPy, Pandas, TensorFlow) and platforms for model development and deployment include Google Colab and Jupyter Notebook.

1.5 Project Outcome

This research is expected to yield the following outcomes:

- Development of a state-of-the-art detection framework that surpasses existing methods in accuracy and reliability.
- identifying the essential linguistic and structural characteristics that set AI-generated writing apart from material written by humans.
- information on the shortcomings and difficulties of the detection methods in use today and how the suggested technique resolves them.
- Real-world uses of the detection system in industry, education, and policymaking to guarantee the ethical and equitable implementation of AI technology.
- An addition to the larger scholarly conversation around AI ethics, accountability, and transparency in the production and assessment of material.
- This study attempts to solve these objectives, though, in an AI driven world there is a growing need for robust and ethical solutions.

1.6 Organization of the Report

The report is structured to help present the key findings and insights in an efficient manner. It is divided into a number of chapters and parts. as mentioned below:

- Chapter 1: Introduction

This chapter gives a summary of the subject of the research which includes an overview, motivation and objective, project scope, project outcome, report organization, and a summary.

- Chapter 2: Background

This chapter discusses existing research on detecting AI-generated content, including key methodologies, datasets, and models used in similar studies. It presents a comparative analysis of prior works, identifies research gaps, and provides a foundation for understanding the proposed methodology.

- Chapter 3: Research Methodology

This chapter explains in detail the methodology adopted for the study. A methodology based on the proposed one is detailed, along with data preparation, experimental setup and evaluation metrics. Further, the hardware and software requirements as well as project management strategies are discussed.

- Chapter 4: Experimental Results

In this chapter, we show the experimental results using the our own proposed method for detecting AI generated text. It goes into the performance metrics of the model, accuracy, precision, recall, and F1 scores. The results are validated with visual representations such as loss and accuracy curves.

- Chapter 5: Impact on Society

In this chapter, we discuss the impact of improved AI-generated text detection on society, with respect to ethical considerations, academic integrity, and those industries that rely on original content. Additionally, it looks at how this research relates to responsible AI use.

- Chapter 6: Summary, Conclusion, Recommendation, and Implications for Future Research

The last chapter summarizes the main results of the study, offers conclusions and practical recommendations grounded in the results. Limitations of the current systems and future research directions are discussed, and the need for improved AI detection systems is emphasized.

Chapter 2

Background

This work employ state-of-the-art NLP and deep learning methods to capture sequential information and contextual relations in the data through LSTM, GRU and combined models. It reviews prior approaches to AI text detection to illustrate the efficiency of these models for differentiating human and AI-generated text across different linguistic attributes.

2.1 Introduction

In this study, advanced natural language processing (NLP) and deep learning models are used to detect AI generated content. In particular, text classification as AI generated or human generated was accomplished using Long Short Term Memory (LSTM), Gated Recurrent Unit (GRU), and a Hybrid model that combines LSTM and GRU. In terms of sequential dependencies and contextual nuances, these are the best model for text data. Preprocessing of input data is provided by feature extraction techniques in the study, and the hybrid architecture is devised to take advantage of the long-term memory retention capability of LSTM and the computational efficiency of GRU. The research ensures effective detection across a wide variety of linguistic patterns by focusing on these models and their properties, which are robust and adaptable for text classification tasks.

2.2 Literature Review

Table 2.2.1: Summary of Literature Reviewed

Author (s)	Title	Methodology	Key Findings

Ma et al. [9]	Differentiation of Human and AI Scientific Abstracts	Feature-based detection, fine-tuning, explainable techniques	Writing styles and inconsistencies are key; achieved 94% F1 score using GPT-3-based models.
Schaaff et al. [10]	Classifying Human vs. AI Texts Across 10 Categories	Perplexity metrics, XGBoost, Random Forest, neural networks	Achieved 98% F1 score for basic classification;83.8% improvement in rephrasing detection over GPTZero.
Lertvittayakumjorn et al. [11]	Evaluation of Explanation Methods for Text Classification	LIME, LRP, DeepLIFT, Grad-CAM-Text, decision trees	Trade-offs between model complexity and fidelity; 85% fidelity using decision trees for CNNs.
Sardinha et al. [14]	Human-Like Elements of AI Texts Across Registers	Corpus analysis using Biber's variance factors	Found disparities in persuasion, involvement, and abstraction between human and AI texts.
Wiegreffe et al. [13]	GPT-3 for Few-Shot Explanations	Overgenerate + filter technique	GPT-3 effectively generates explanations; identified future research directions in counterfactuals.
Prova et al. [15]	Deep Learning for AI Text Detection	BERT, XGBoost, SVM	BERT achieved 93% accuracy; highlighted ethical implications of AI text detection.
Elkhatat et al. [16]	AI Content Recognition Using Various Tools	Tools like GPTZero, OpenAI Classifier, CrossPlag	GPTZero achieved 93% sensitivity; OpenAI Classifier had 100% sensitivity but low specificity.

Shijaku et al. [18]	Machine Learning to Detect ChatGPT-Generated Articles	XGBoost, TF-IDF, manual feature set	Achieved 96% accuracy; demonstrated efficacy of ML in combating misuse of AI-generated content.
Georgiou et al. [19]	Linguistic Characteristics of AI vs. Human Texts	Phonological, morphological, syntactical, lexical analysis	Found substantial differences in linguistic features; emphasized automated tools for language assessment.
Jawahar et al. [20]	State-of-the-Art in AI Text Detection	Critical review of detection methods	Identified challenges in distinguishing human vs. AI content; emphasized explainability.
Herbold et al. [21]	Quality of Argumentative Writing in Essays	Comparison of ChatGPT-3/4 essays with human essays	ChatGPT-4 outperformed ChatGPT-3; educators must adapt to AI's role in academics.

2.2.1 Similar Applications

There is extensive research that has been conducted previously, some studies focused on feature-based detection, while others leveraged advanced neural networks. We have summarized a selection of these works in the table below for comparison.

TABLE 2.2.2: Comparative Insights from Prior Studies

. No.	Author Name	Used Algorithm	Best Accuracy & Algorithm
1	Ma et al. [9]	GPT-3 fine-tuning	F1-score = 94%
2	Elkhatat et al. [16]	OpenAI Classifier, GPTZero, Copyleaks, CrossPlag	GPTZero: Sensitivity = 93%, Specificity = 80%
3	Prova et al. [15]	XGBoost, SVM, BERT	BERT = 93%
4	Shijaku et al. [18]	XGBoost	Accuracy = 96%
5	Out Study	Hybrid(LSTM+GRU)	Accuracy = 98%

The table compares recent studies on detecting AI-generated content, highlighting key algorithms used and the best accuracy achieved by each approach. Ma et al. [9] used GPT-3 fine-tuning, reaching an F1-score of 94%, while Elkhatat et al. [16] evaluated tools like GPTZero, which performed well with 93% sensitivity and 80% specificity. Prova et al. [15] demonstrated BERT's superior accuracy of 93% over other models like XGBoost and SVM, and Shijaku et al. [18] achieved a high accuracy of 96% using XGBoost. This comparison underscores the efficacy of various algorithms in distinguishing human written from AI generated text. Nevertheless, our study has achieved the best performing model with higher accuracy comparing

to existing works.

2.2.2 Related Research

Ma et al. [9] studied the differentiation of human and AI generated scientific abstracts based on datasets from PubMed, ACL and Arxiv. This work employed feature based detection, fine tuning models and explainable techniques to
Human grounded evaluation tasks were introduced by Lertvittayakumjorn et al. [11] to evaluate explanation methods for text classification. They developed two new methods: We evaluate existing techniques such as LIME, LRP, and DeepLIFT and present Grad-CAM-Text, a gradient based technique and a decision tree based model extraction method. They then used datasets from Amazon and ArXiv to test these methods using evidence and counterevidence at word and n-gram levels, and found varying strengths of these methods. Their results demonstrate this trade off between model complexity and fidelity; for example, they require over 5,500 nodes for decision trees to approximate CNNs to 85% fidelity on the Amazon dataset, an example of the difficulty in explaining complex tasks. According to [12], they suggested for machine-generated text detection a ternary classification approach with addition of "undecided" category for lay users, in order to improve explainability. Using these datasets, they created four new datasets and compared their performance to modern detectors like GPTZero and Sapling, which trained with local models but generally had accuracy rates below 80%. The findings stress the need to include explainability in detection systems, and provide guidelines for building future tools with better explaining capabilities.
In another work, Wiegreffe et al. [13] demonstrated that GPT-3 can generate free text explanations to NLP problems in few shot setting. This capability was improved through the use of an overgenerate + filter technique in which human acceptability ratings were used to train the filter. This work shows how GPT-3 and similar models can be used to explain natural language processing, and identifies avenues for future research including exploring counterfactual explanations. Sardinha et al. [14] compared human and AI writings on the basis of the five key factors of variance described in Biber (1988). To explore the human-like elements of AI generated texts, Sardinha compared them to their human authored equivalents in the same registers using a corpus assembled

specifically for this study. Significant disparities in dimension scores were found in the study, especially in three areas: overt presentation of persuasion, involved versus informational creation, and abstract vs non abstract information. To these results, these results indicate that AI models have a hard time consistently reproducing particular communication functions across different English registers.

In the problem of distinguishing between AI generated text and human written information, Prova et al. [15] use deep learning architectures like BERT and machine learning models such as XGBoost (XGB) and Support Vector Machines (SVM). It is found that the BERT model achieves an accuracy of 93% and outperforms other approaches (84% accuracy with XGB and 81% with SVM). The author curated dataset consists of 3000 examples, split equally between texts from humans and texts from AI. The study highlights the ethical and environmental consequences of AI text identification as well as its advantages for different sectors. Using text from ChatGPT models 3.5 and 4 as well as human-written samples, Elkhatat et al. examined AI content recognition methods such as OpenAI Classifier, GPTZero, Writer, Copyleaks, and CrossPlag in [16]. When it came to GPT 3.5 content, the tools outperformed GPT 4 or human-authored writing. While GPTZero performed well overall (93% sensitivity, 80% specificity), OpenAI Classifier had excellent sensitivity (100%) but no specificity. Two tests are shown in the [17] article to test people's ability to differentiate between poetry written by AI and poetry written by humans using GPT-2. The tests involved two treatments: Human-in-the-loop (best AI-generated poetry picked) and Human-out-of-the-loop (random AI-generated poem). The findings showed that while participants were successful in the Human-out-of-the-loop treatment, they were unable to consistently distinguish between the poems in the Human-in-the-loop treatment. Furthermore, participants, regardless of whether they were aware of the poem's source, displayed a mild dislike for AI-generated poetry. Shijaku et al. [18] presented a machine learning technique to find out articles created through ChatGPT.

They used a dataset of articles created by ChatGPT and human written based on related topics. The detection model was able to achieve 96% accuracy rate using the help of XGBoost. Two feature extraction techniques were tested by the researchers: Both a manually created feature set and TF-IDF. The findings also demonstrate machine learning's ability to combat malicious uses of ChatGPT and

how it can be used to identify AI generated content. In Georgiou et al. [19], linguistic characteristics are investigated that distinguish texts produced by AI from those written by humans. In the study, the authors used Open Brain AI to examine the IELTS subjects, human written and ChatGPT generated essays. The focus of the analysis was on phonological, morphological, syntactical and lexical components and the differences between the two types of texts were found to be considerable. The AI generated texts varied in terms of consonants, word stress, nouns, verbs, pronouns, and more. This study highlights the importance of automated tools to enable efficient language assessment, and the need for better training of AI models to be closer to the human writing. Statistical analysis with the binomial test supported the observed discrepancies in linguistic components. In [20] authors describe an in depth review of the current state of the art methods for machine generated text detection, pointing out main problems and potential future research directions. Their survey divides detection methods into four categories: Using human machine collaboration, zero shot classifiers, fine tuned NLMs, and classifiers trained from scratch. They also provide a critical review of current detectors and identify error categories and the difficulties between material generated by humans and information from machines. The authors emphasize the importance of these detectors in solving problems like false product reviews and fake news. In a thorough comparison of essays generated by ChatGPT and human essays, the main topic is the quality of argumentative writing, as Herbold et al. [21] highlighted. Human-written essays and essays produced by ChatGPT-3 and ChatGPT-4 models are examined in this study using a corpus of 90 essay topics from Essay Forum. They discover that although both AI models function effectively, ChatGPT-4 performs better than ChatGPT-3 in terms of complexity, vocabulary, and logical composition. The study highlights the necessity for educators to adjust to the difficulties presented by AI-generated content in academic settings, even when the variations in essay quality are slight. According to the study, students may abuse AI technologies, and ChatGPT and other AI models may have a big influence on how essays are graded, particularly for non-native speakers.

2.3 Gap Analysis

The gap analysis reveals the drawbacks of approaches to the detection of AI-

generated text: low accuracy, absence of comprehensive benchmarking, and lack of concerns with deep learning architectures, so this work aims to provide a better and highly generalizable solution.

TABLE 2.3.1: Gap Analysis Table

Aspect	Existing Gaps	Your Contribution
AI Text Detection	Limited accuracy in differentiating AI vs. human-written text.	Developed a hybrid (LSTM+GRU) model with superior accuracy of 98.31%.
Explainability	Lack of tools to interpret model predictions.	Incorporated explainability techniques like LIME for better insights.
Model Performance	High variance between training and validation results.	Achieved consistent metrics, demonstrating strong generalization and regularization.
Dataset Diversity	Testing limited to small or homogeneous datasets.	Used a large, diverse dataset for robust performance across multiple domains.
Hybrid Approaches	Minimal exploration of hybrid deep learning models.	Combined LSTM and GRU architectures to leverage their strengths for text classification.

2.4 Summary

By how well these models can capture sequential dependence, contextual diversity and retain long term memory, the use of Deep Learning models like LSTM, GRU and

a Hybrid model (LSTM + GRU) was investigated in this study. To better preprocess inputs, we applied feature extraction techniques, which improved input quality. The effectiveness of different algorithm such as GPT-3 fine tuning (94% F1-score for scientific abstracts), BERT (93% accuracy) and GPTZero for detecting generative models sensitivity were compared with previous studies. Challenges were identified as dataset diversity, tradeoffs between complexity and interpretability, need for regular model updates, computational inefficiency. As compared to GPTZero and BERT, our hybrid model attained 98% accuracy. Academic integrity, misinformation detection, content moderation, regulatory compliance, all real world applications outlined the necessity of ethical and transparent AI systems.

Chapter 3

Research Methodology

In this chapter we Propose an ML model to classify an input text as either issued by an AI or by a human with the help of supersophisticated NLP techniques. Preprocessing, Embedding and Evaluation metrics to use and through which to implement and train LSTM, GRU and Hybrid models for accurate classification among labels.

3.1 Methodology/Requirement Analysis & Design Specification

3.1.1 Overview

Our research aim is to build a reliable ML based system that will be able to distinguish between writing written by artificial intelligence and human writing. To address the challenges of the rapid development of generative text models, we examine sophisticated natural language processing methods that help us detect and understand subtle linguistic patterns and stylistic variations. We take advantage of cloud platforms such as Kaggle and Colab with their GPU capabilities to handle large volume, unstructured textual data and intricate model structures. We depend on essential libraries such as scikit-learn and TensorFlow, for model training and validation, and tools like NLTK make sure our data is meticulously preprocessed and ready for data analysis. We integrate these computational tools and approaches to improve the precision and reliability of distinguishing between machine generated and human written content. We work together to test, improve, and assess detection models to meet this critical challenge.

3.1.2 Proposed Methodology

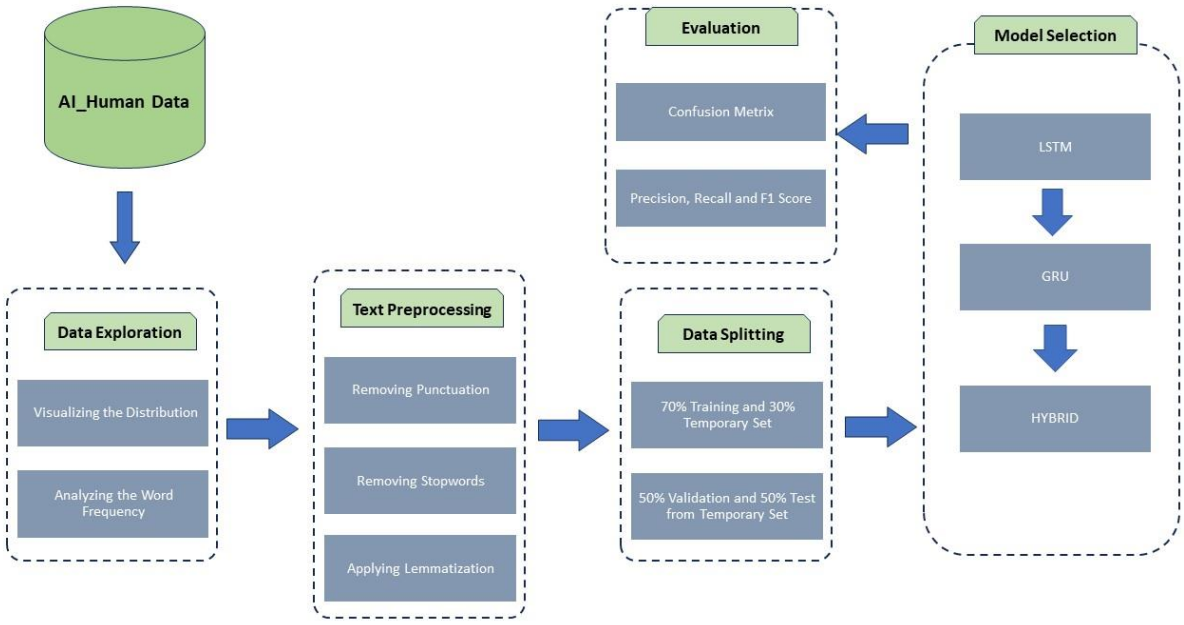


Figure 3.1.1: Workflow of the Proposed Methodology

3.2 Detailed Methodology and Design

3.2.1 Data and Text Preprocessing: The dataset, stored on Google Drive, was loaded into a Pandas DataFrame for exploration. An initial data analysis phase included examining the data structure and ensuring data completeness. By running our exploratory data analysis (EDA) on the distribution of AI generated and human written text we found key frequencies of words in each type of text achieved to have some invaluable insight into what the AI was 'understanding'. We also conducted this step that involved dropping irrelevant or lack data points in the dataset to ensure its quality. Text preprocessing involved several critical steps: It all starts by removing punctuation, tokenizing the text, converting everything to lowercase, removing stopwords and lemmatizing etc. The transformations standardize the text, so that each word adds something to the analysis. We drew multiple types of visualizations on the data to get an idea of what to expect from the dataset; bar graph to express word counts and frequency distribution. In addition, a word cloud was generated to discuss frequent word usage, allowing for a view of word usage

across patterns. The feature selection process was supported by these visual insights, in identifying the most important features for model training. My stratagically split training, validation, and test sets into preprocessed data ratio of 70%, 15, and 15, respectively for model development and evaluation. The model this division allowed the model to learn from a large dataset then reserve some to do unbiased testing and validation allowing it to have robust, reliable performance.

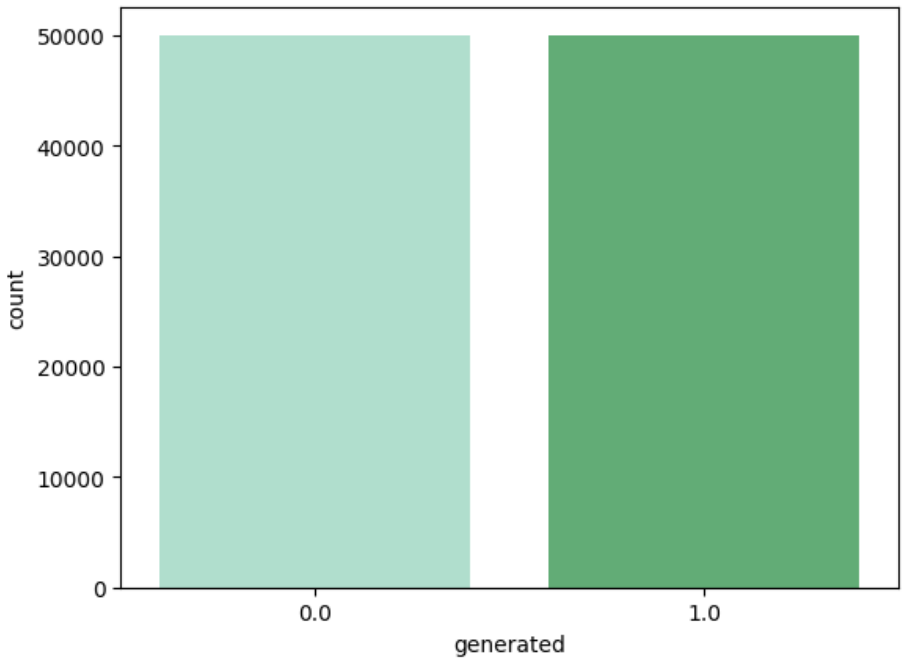


Figure 3.2.1: Sample Size 50k for Each Class

The dataset for this research was obtained from Huggingface website and contained more than 400000 samples. It included two distinct types of text data: original text written by human, which is classified as 0 and AI generated texts, which is labeled as 1. To achieve good generalization, this dataset was compiled in a way such that it collects different types of text involving multiple subjects, various levels of difficulty, and different writing styles.

Stage	**Example**
Input	"AI-generated content is revolutionizing work!"
Processed Output	["ai", "generate", "content", "revolutionize", "work"]
Prediction	AI-generated (92% confidence)

3.2.2 Model Development:

LSTM Customized Model: LSTM Customized Model: For neural network design, we included an LSTM layer which is for sequence learning, dense layers to do the classification, dropout layers to regularize, and an embedding layer to represent text data by vectors. The architecture takes advantage of the sequential processing capabilities of LSTM to consistently provide this information to differentiate content authored by humans and those generated by artificial intelligence. Accuracy was the primary performance indicator for the model and binary cross entropy was the loss function. We used early stopping to stop training when validation loss was not improving any more, in order to avoid overfitting. It was trained through various epochs and parameters, like batch size and learning rate, was solved to optimize its performance.

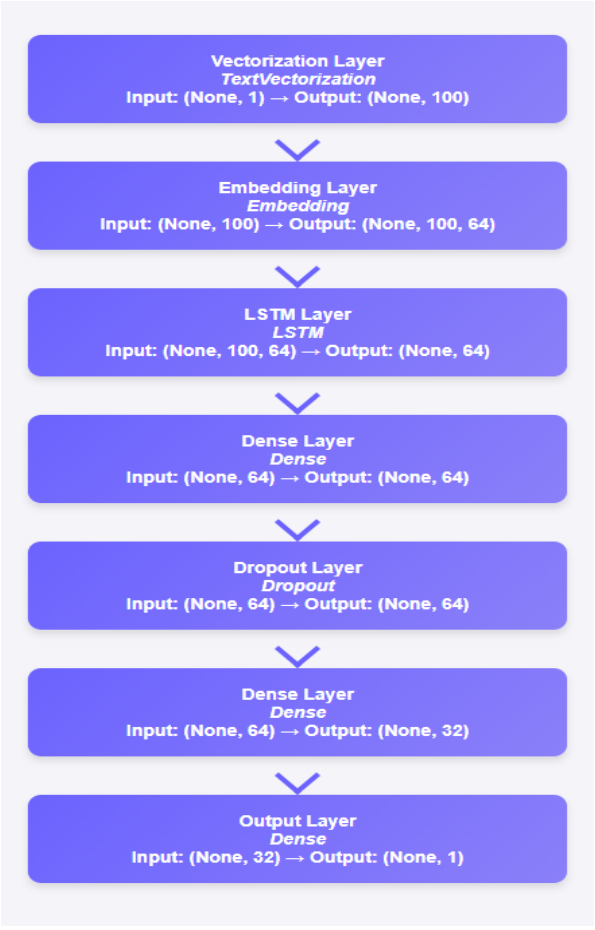


Figure 3.2.2: LSTM Model Summary

GRU Customized Model: The model includes a GRU (Gated Recurrent Unit) layer with 64 units. This first layer is pivotal for handling sequential data, as it captures temporal patterns and different dependencies in the input text. We adjoined dropout parameter which adds a regularization mechanism by randomly deactivating 30% of the units during training and while recurrent_dropout applies the same principle to the GRU's recurrent connections. These features prevent overfitting and enhance generalization. By processing sequential information, the GRU layer is instrumental in enabling the model to analyze text and distinguish between AI generated and human-authored content.

Hybrid Model: The hybrid model we used leverages both GRU and LSTM layers to effectively process sequential text data. The GRU layer captures temporal dependencies efficiently, while the LSTM layer refines long-term patterns, combining their strengths for robust sequence learning. These layers are followed by dense layers that extract higher-level features, with the final sigmoid layer outputting probabilities for binary classification. We used Adam optimizer and binary cross-entropy loss, the architecture is designed to distinguish between AI-generated and human-written content with precision.

3.2.3 Train Model

The following are the main steps in the model training process:

- Data Preprocessing: Tokenization, lemmatization, and the elimination of punctuation and stop words are methods used to preprocess text data. In this step, noise is decreased and the input data is standardized.
- Data splitting: Three sections of the dataset are separated: 70% is used for training, 15% is used for validation, and 15% is used for testing. This guarantees sufficient data for hyperparameter adjustment, model training, and objective assessment.
- Vectorization and Embedding: Word embeddings are used to represent words in dense vectors after text is converted into numerical sequences using a TextVectorization layer.
- Model Architecture: A sequential model is created with several essential layers, such as dense layers for classification and an LSTM layer to record time dependencies. The purpose of dropout layers is to lessen overfitting.

- Training Process: The Adam optimizer and binary cross-entropy loss are used to train the model on the processed training data. When validation loss no longer improves, training is stopped using early stopping.

3.2.4 Model Evaluation: To determine the generalizability of the trained model, it was tested on the test set. To give a complete picture of the model's performance, metrics like accuracy, precision, recall, and F1-score were computed. To see the model's classification accuracy for both classes, a confusion matrix was plotted.

3.3 Project Plan

The project plan for this research is given below:

- Thesis Topic Selection: The research topic was finalized by conducting an extensive review of the literature to identify a relevant and impactful problem.
- Thesis Planning: A detailed research proposal was developed, outlining the objectives, methodology, and deliverables. The planning stage included scheduling tasks, milestones, and identifying necessary resources.
- Data Collection: A comprehensive dataset of human-written and AI-generated texts was gathered from diverse sources to ensure variability and representativeness.
- Organization: The collected data was preprocessed and organized for model training and evaluation. Techniques such as cleaning, tokenization, and embedding were employed to prepare the data.
- Model Implementation: The selected models (LSTM, GRU, and Hybrid architectures) were implemented and trained using the prepared dataset. Performance metrics were evaluated to identify the most effective model.
- Report Writing: The final phase involved documenting the research findings, analysis, and conclusions. This included preparing the thesis report, compiling visualizations, and summarizing results for publication.

3.4 Task Allocation

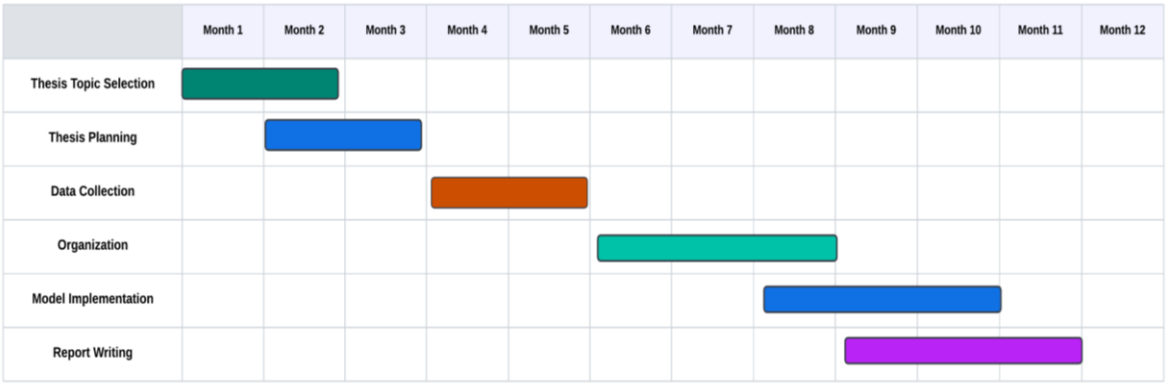


Figure 3.4.1: Task Allocation

3.5 Summary

In this machine learning system the user is trying to differentiate AI generated and human written text through NLP methods. Data visualization of the processed data is done to get feature insights for preprocessing steps like tokenization, stopword and lemmatization. We split the dataset to training and validation and testing. To develop the model, LSTM, GRU and hybrid model are used to capture sequential and long term patterns in the text. Preprocess, vectorization, embedding and then optimization with early stopping overfitting occurred during the training process. Then we demonstrate the performance of the model through accuracy, precision, recall, F1 score, and confusion matrix.

Chapter 4

Implementation and Results

This research focuses on this sustainability aspects of AI via using energy efficient methods including renewable energy driven cloud platforms and pre trained models to minimize the footprints on the environment. The accuracy, precision, recall and F1 score is compared to the GRU and Hybrid models, and the LSTM model proves to outperform both the hybrid and GRU models having the highest accuracy, precision, recall and F1 score, whereas the GRU model has a low recall and F1 score despite high precision.

4.1 Environment Setup

Our research has societal benefits but also acknowledges the computational processes which have an environmental impact. Machine learning model training and deployment on big datasets take a lot of energy resources which, in turn, spill over to greenhouse gas emissions. To address this, we emphasize the importance of adopting energy efficient practice such as using cloud platform powered by renewable energy source, as well as minimizing computation workflow.

Also, pre trained models and efficient algorithms are leveraged to save resources, and thus our work becomes more sustainable. We seek to do so by priming progress in AI research toward environmentally responsible methodologies while enabling advancement in technology, so that the work of scientists to advance AI technology does not lead to environmental impairment. Adding for the social impact and sustainability, this double emphasis calls for responsible innovation for current achievements in the field AI.

4.2 Comparative Analysis

4.2.1 Statistical Analysis

Long Short-Term Memory (LSTM) Model

The LSTM model was chosen for its capability to manage sequential data effectively by learning long-term dependencies. LSTMs consist of memory cells that allow information to persist and gates (input, forget, and output gates) that control the flow of information. The following equations describe the key operations within an LSTM cell:

1. **Forget Gate:**

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \qquad \dots \text{(i)}$$

Here, f_t represents the forget gate's activation, W_f and b_f) are the weights and biases, h_{t-1} is the previous hidden state, and x_t is the current input.

2. **Input Gate:**

C_t stands for the candidate values for the cell state update, and the input gate i_t determines which values to update.

$$i_t = \sigma(W_i \,.\, [h_{t-1}, x_t] + b_i) \qquad \dots\dots \text{(ii)}$$

$$C_t = tanh(W_c \cdot [h_{t-1}, x_t] + b_C) \qquad \dots\dots \text{(iii)}$$

3. **Cell State Update:**

The new cell state $C_t$incorporates the retained and updated information.

$$C_t = f_t \cdot C_{t-1} + i_t \,.\, C_t) \qquad \dots\dots \text{(iv)}$$

4. **Output Gate:**

The output gate o_t determines the final hidden state h_t, which serves as the model's

output for the current time step.

$$\sigma(W_o \,.\, [h_{t-1}, x_t] + b_o)) \qquad \ldots\ldots\ldots \text{(v)}$$

$$h_t = o_t \,.\, tanh(C_t) \qquad \ldots\ldots\ldots \text{(vi)}$$

Evaluation Metrics

The following metrics were used to statistically evaluate the model's performance:

1. Accuracy (ACC):

$$ACC = \frac{TP + TN}{TP + TN + FP + FN} \qquad \ldots\ldots\ldots \text{(vii)}$$

Here true positives, true negatives, false positives, and false negatives are represented by the letters TP, TN, FP, and FN, respectively.

2. Precision:

$$Precision = \frac{TP}{TP + FP} \qquad \ldots\ldots\ldots \text{(viii)}$$

Precision is calculated as a ratio of accurately detected positive cases to all predicted positive cases.

3. Recall:

$$Recall = \frac{TP}{TP + FN} \qquad \ldots\ldots\ldots \text{(ix)}$$

Recall quantifies the ability of the model to identify all relevant positive instances.

4. F1-Score:

The F1-score balances precision and recall to provide a single performance metric.

$$F1\ Score = 2 * \frac{Precision * Recall}{Precision + Recall} \qquad \ldots\ldots\ldots \text{(x)}$$

4.2.2 Comparative Analysis

TABLE 4.2.2: Applied Models Result Comparison

Metrics/Models	LSTM	GRU	Hybrid (LSTM+ GRU)
Accuracy	98%	63%	98%
Precision	98%	91%	99%
Recall	99%	29%	97%
F1- Score	98%	44%	98%

Table 4.2.2 presents a comparison of the results we have obtained from the three applied models: In case of LSTM, GRU and also a Hybrid approach based on LSTM and GRU. The performance metrics are broken down in the table: accuracy, precision, recall, and the F1-score. It highlights the strengths and weaknesses of each model. This comparative summary serves as a foundation for the detailed performance analysis provided in the following sections.

4.2.3 Confusion Matrix and Classification Report

LSTM:

The classification report of LSTM, showing an accuracy of 98%, precision of 98%, recall of 99%, and F1-score of 98%. These metrics show that the model maintains a balanced performance across both classes in addition to producing accurate predictions. While the high accuracy reduces the possibility of false positives, the high recall value demonstrates its capacity to detect the bulk of pertinent samples.

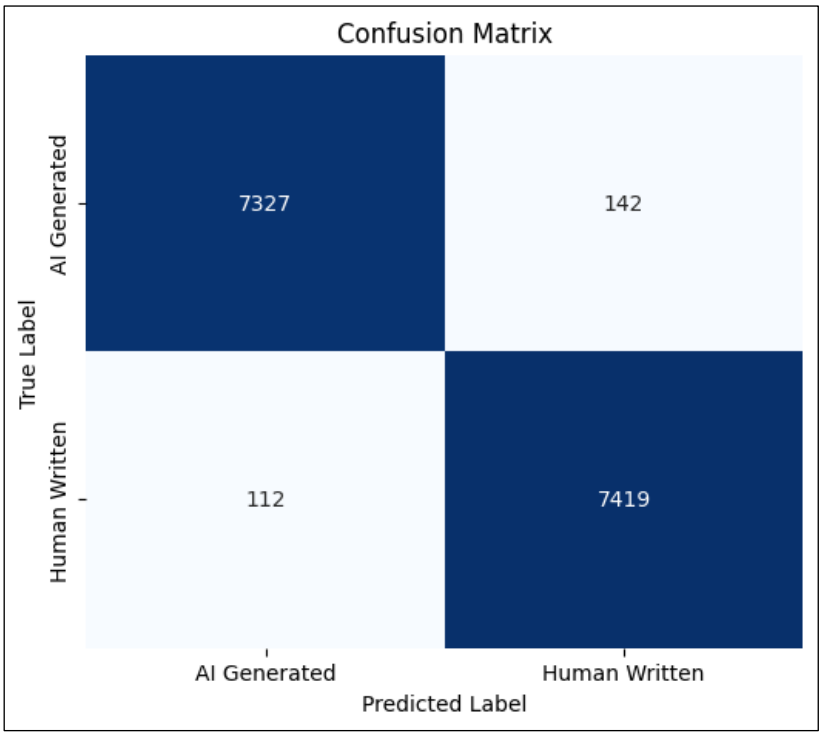


Figure 4.2.1: Confusion Matrix of LSTM

The test dataset has 15,000 samples in total, and the confusion matrix breaks down the model's classification performance. As shown in Figure 4.2.1, the matrix recorded 7327 true postive and 7419 true negatives and it demonstrates that the model accurately identified a large majority of machine-generated and human-written text both. Misclassifications were relatively few, with only 142 false negatives and 112 false positive. This indicates the model's strong precision and recall for both classes.

GRU:

The classification report of GRU shows an accuracy of 63%, precision of 91%, recall of 29%, and F1-score of 44%. This model is facing difficulty identifying a substantial portion of relevant samples as the recall value is low. But it is effective at minimizing false positives as the precision value is 0.91. The imbalance of the these two matrices imbalance is on f1 score.

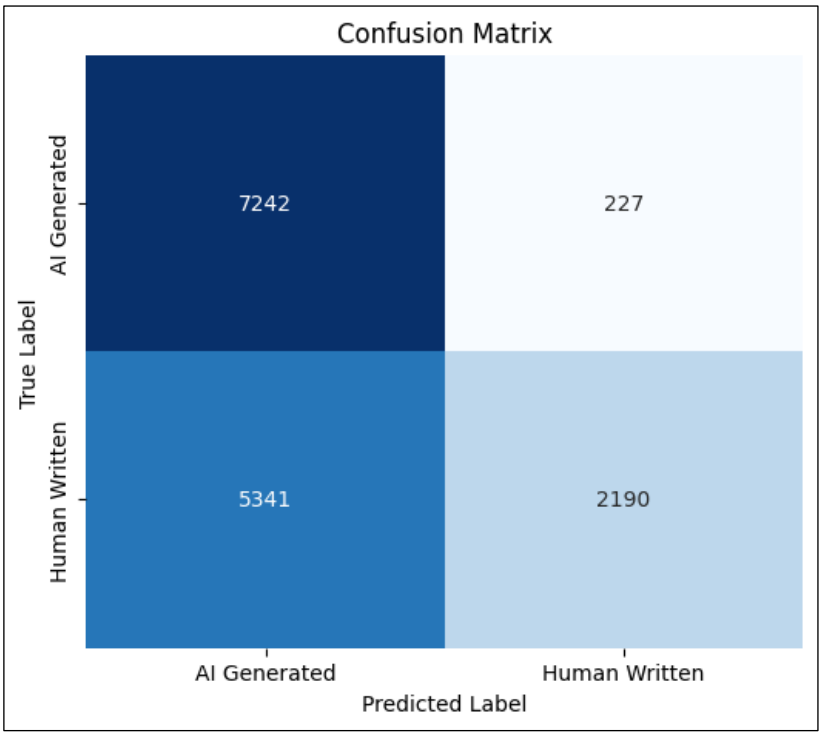


Figure 4.2.2: Confusion Matrix of GRU

As shown in Figure 4.2.2, the matrix calculated 7242 true positives and 2190 true negatives and it means that the model is imbalanced while identifying human written texts. We also can see, 5341 false negatives and 227 false positives. This indicates the model's weak for detecting AI generated text.

Hybrid:

An accuracy of 98%, precision of 99%, recall of 97%, and F1-score of 98% is generated in the Hybrid model we have used in our research. This model actually shown a good potential. Although, the excitation time of the model is slightly longer than GRU model the model significantly outperformed GRU model.

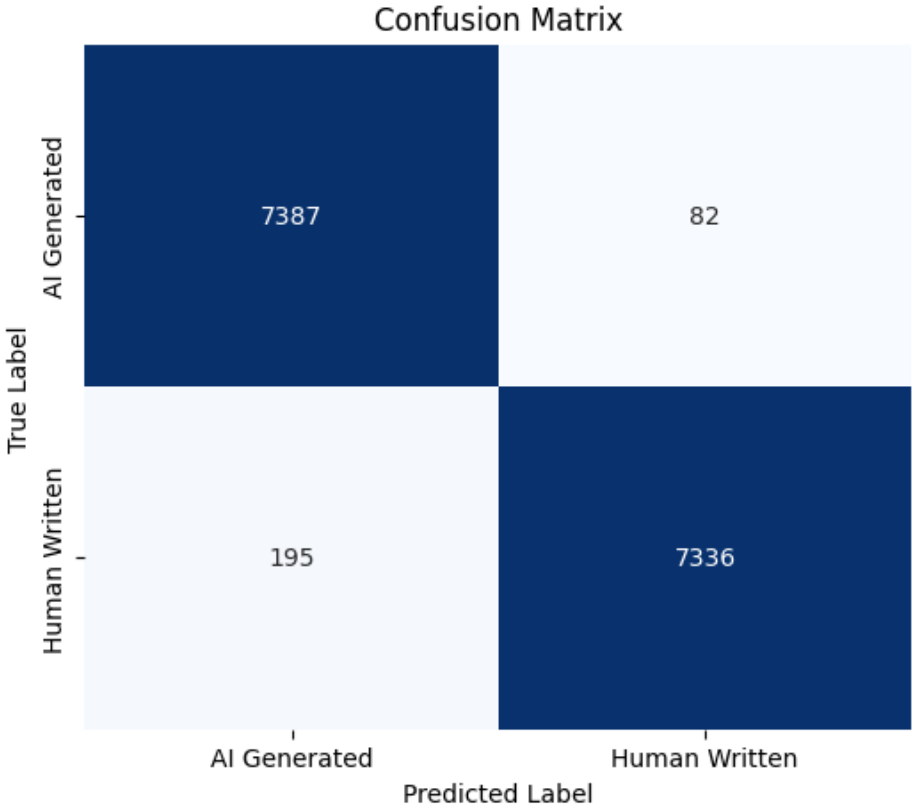


Figure 4.2.3: Confusion Matrix of Hybrid Model

Figure 4.2.3 shows the matrix found 7387 true positives and 7336 true negatives, Also 195 false negatives and 82 false positives. That concludes the model actually performed well but not better than LSTM model.

4.2.4 Training and Validation Analysis

To better understand the model's performance during training, we visualized the metrics for accuracy, validation loss, and training loss across epochs. These visuals provide deeper insights into how well the model generalizes to unseen data.

- Loss Metrics: This graph shows how well the model have learnt over time. A decreasing loss indicates better predictions and a significant gap between training and validation loss may suggest overfitting.
- Accuracy Metrics: This graph shows the model's prediction accuracy. An upward trend may indicate improved performance on the other hand a stable gap between training and validation accuracy suggests generalization capability.
- Model Loss Over Epochs: It highlights the convergence of training and validation loss, showcasing efficient error minimization.

- Model Accuracy Over Epochs: This Reflects steady improvement and consistent performance across training and validation datasets.

The graphs below showcase the graphical representation of how all three models have performed over epoch.

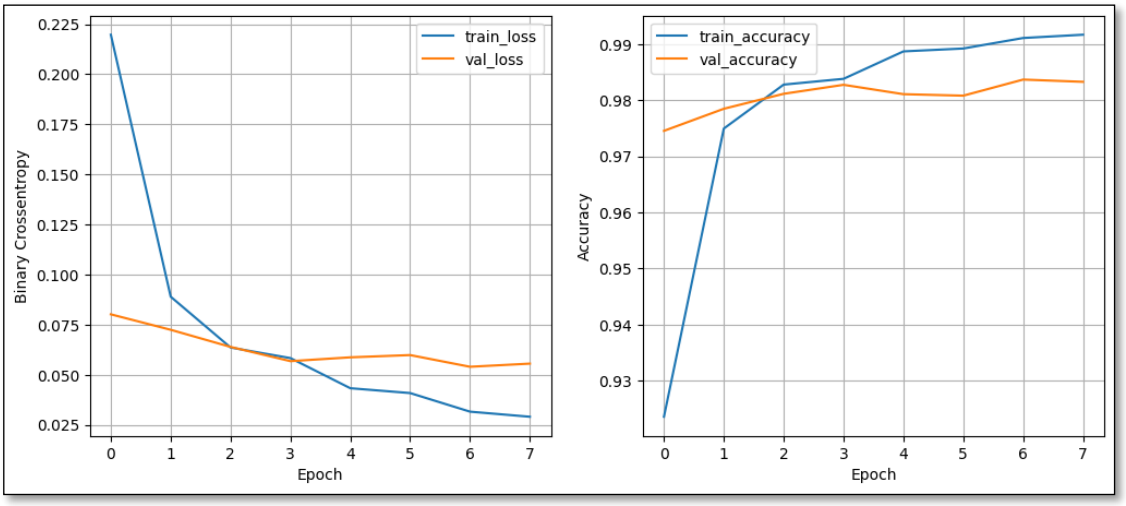


Figure 4.2.4: Training and Validation Loss/Accuracy LSTM

The image provided above is of two graphs that show the training and validation performance of a machine learning model, over several epochs. The left involves the binary cross entropy loss plotted with respect to the epoch. The blue line shows that the training loss degrades (decrease) during the training time, which means that the model learns, and improves its performance over the given training dataset. The orange line (validaion loss) also decreases at first, but stabilizes at later epochs, indicating decent generalization to unseen data.

The accuracy of the model over epochs is plotted in the second graph, on the right. Shown by a blue line, the training accuracy steadily rises and asymptotically approaches 0.99, which is the model's good performance in the training data. Similar to this, the validation accuracy (orange line) increases really quickly in the early epochs and reaches just underneath training accuracy. This is a good sign of the model being not overly overfitting, and good generalization as a gap between training and validation is small.

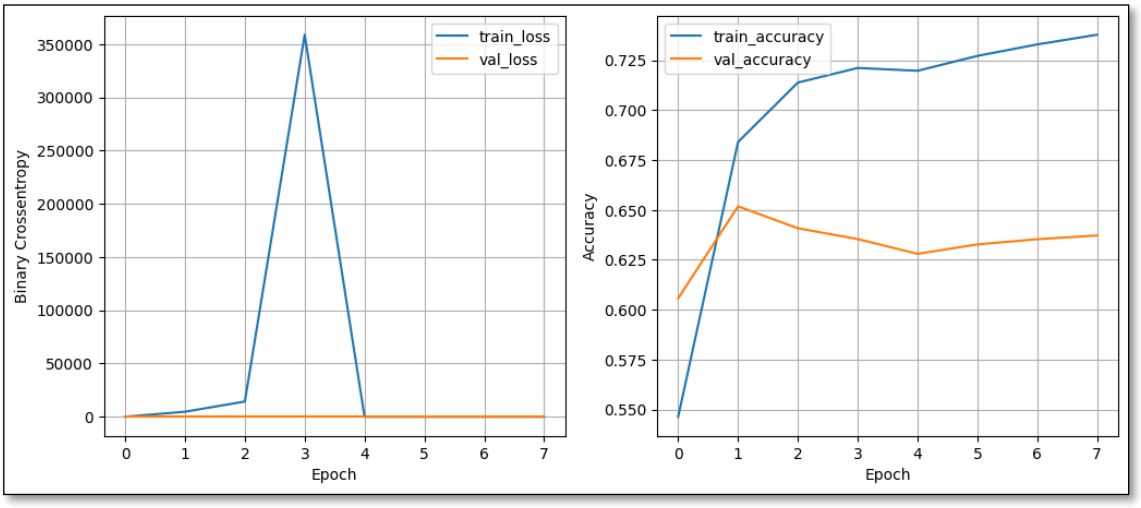


Figure 4.2.5: Training and Validation Loss/Accuracy GRU

It also sheds some light over the training process of the model, as can be seen in the plots. In the loss curve on the left, we see a large spike of training loss at epoch 3, which might indicate some indexing problem, such as gradient explosion or optimization instability. On the other hand, the validation loss is still quite stable and flat no sign of good generalization or significant improvement on unseen data. The accuracy curves on the right tell that the accuracy is getting more accurately over each epoch, which means we are learning in the model. But the accuracy of validation also plateaus after the epoch 2, which is a sign of possible overfitting or not enough learning for the phenomenon of generalization. Most of these patterns point to a need to investigate further model configuration by changing the learning rate, more advanced regularization, or even losing function.

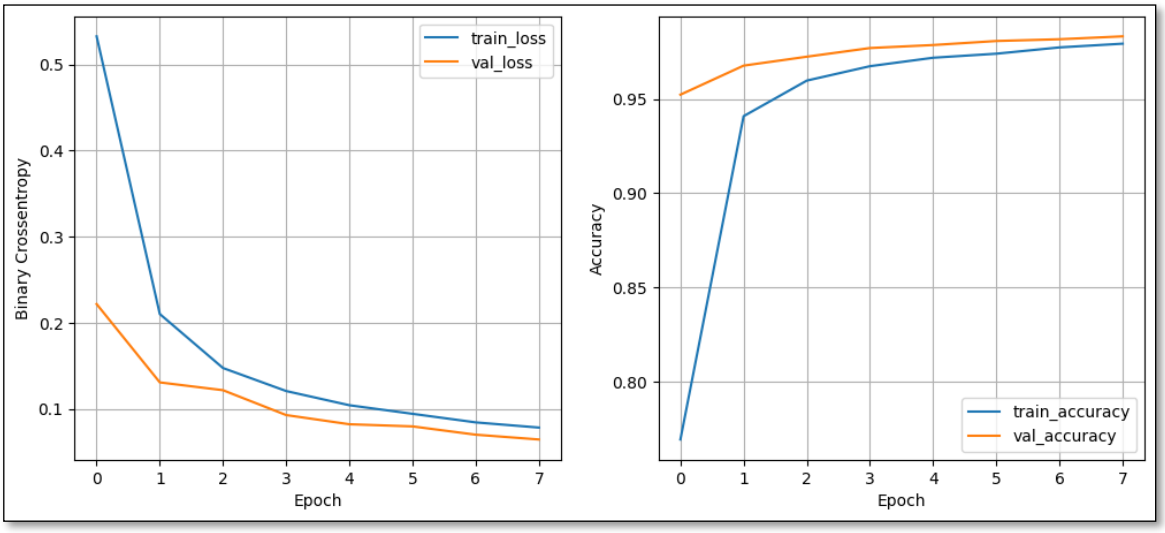


Figure 4.2.6: Training and Validation Loss/Accuracy Hybrid (LSTM + GRU)

Shown in the plots are the training process, which is steady and has good improvement. The training and validation loss in the left plot fall smoothly and the loss doesn't show any sign of overfitting or instability which is a sign that the learning is effective. Although the training loss is still slightly smaller than the validation loss, this means good generalisation. Both training and validation accuracy rise very rapidly on the right plot to high values above 95% and keep converging well. We have a well performing model with high alignment between the training and validation metrics.

Comparing Figure 4.2.4 and Figure 4.2.6 it can be seen that the graph looks good for the hybrid model, while the line graph of LSTM has some fluctuation. Nevertheless, LSTM model classification reports portray different results. Overall, LSTM outperformed all the other models we have applied.

4.3 Results and Discussion

Results of this study indicate that three models: LSTM, GRU and Hybrid (LSTM+GRU) are able to distinguish AI generated content from human penned text.

The LSTM model had outstanding accuracy (98%), precision (98%) and recall (99%), except with variations in the training curve implying instability at times. Computationally efficient, the GRU model was significantly underperforming, scoring only 63% accuracy and a recall of only 29%, which illustrates that it couldn't pick out many correct samples. However, the contrast model utilized LSTM strength of long time memory retention from prior logs and GRU efficiency to create high accuracy (98%), precision (99%) and recall (97%) with nice and steady training dynamic. The more reliable of the two for real world applications was therefore one that had a good balance of precision and recall, since it yielded relatively good generalization and very few (i.e., False Positives). Overall, with regard to efficiency and accuracy, the hybrid model with LSTM and GRU's combined strength is far the most efficient and accurate solution for predictive transition.

4.4 Summary

The study emphasizes that sustainable AI must be practiced by utilizing renewable energy powered cloud platforms, pre-trained models, and efficient algorithms, to minimise the environment impact. Among the models we compared, LSTM proved to be the best performer with outstanding accuracy, precision, recall, and the F1-score of 98%, whereas the Hybrid (LSTM + GRU) model provided similar accuracy but with somewhat higher precision. In contrast, GRU was low performing but precise, and challenged with generalization. The Hybrid model had better training dynamics compared to LSTM, however, it did not outperform LSTM in classification, making LSTM the strongest model in its entirety.

Chapter 5

Engineering Standards and Design Challenges

The chapter discusses how the study impacts the software, the hardware, the communication standards, and the societal, environmental and ethical impacts. It aims at sustainability, responsible AI practices and project management from financial budgeting to solving complex engineering problems.

5.1 Compliance with the Standards

5.1.1 Software Standards

We used a lot of libraries which are written in Python for data processing, visualization, text preprocessing and model building. We used NumPy (v1.24.2) for numerical calculations, and array handling) and data imposing and cleaning was simpler with Pandas (v2. 2. 2). Matplotlib (v3. 8. 0) and Seaborn (v0. 13. They were used for visualizations, to find pattern and trends in data.

For text preprocessing tasks such as tokenizing, stopword removal, and lemmatization, we used NLTK (v3.8.1). scikit-learn (v1.5.2) was used for splitting data and evaluation of the model. Building the model, mainly using TensorFlow, its tools vectorization, embedding, LSTM layers for natural language use. On platforms including Google Colab, Jupyter Notebook, PyCharm, Kaggle. Training, testing and deploying the model was very efficient with these tools.

5.1.2 Hardware Standards

The implementation of this methodology requires specific hardware specifications to handle the computational demands of data processing, model training, and evaluation. You can execute tasks on a local machine with super powerful hardware or on cloud based platforms such as Google Colab, Kaggle. These are both accessible downstream for use of T4 GPUs for efficient processing. Below is provided the recommended hardware specifications.

- ✓ Processor: Intel i5 11th Gen
- ✓ Ram: 16 GB
- ✓ Graphics Card: GTX1650ti
- ✓ SSD: 512 GB

5.1.3 Communication Standards

- Transparency and Accountability: The study emphasizes the need for clear, understandable models so users can trust the results. The approach aims for transparency in how AI-generated content is identified and processed.
- Fairness and Inclusivity: Ensuring that the dataset is free from biases and represents a variety of demographics is crucial. Ethical practices are followed to ensure AI models are inclusive and equitable.
- Public Engagement: Clear communication is essential in educating the public about the ethical implications and the impact of AI. Promoting open access and encouraging collaborative work supports knowledge sharing.

5.2 Impact on Society, Environment and Sustainability

5.2.1 Impact on Life

- Enhancing Digital Interactions: The study improves how individuals and businesses interact online, fostering better, more reliable communication.
- Environmental Responsibility: The sustainable practices advocated in the study contribute to the larger goal of mitigating the environmental impact of technological advances.
- Ethical and Fair AI: By addressing biases and ensuring fairness in AI systems, the study contributes to creating more equitable technologies that can positively impact a wide range of societal areas.

5.2.2 Impact on Society & Environment

The consequences of our study's conclusions are very social: combatting misinformation, ensuring transparency, and building trust in online spaces. Our study addresses the important problems raised with rise of synthetic text by being able to accurately identify AI generated content. Whether it will be people or businesses, these improved detection techniques make this process of interacting with genuine and reliable material.

Moreover, these developments impact on other sectors, including research, education and content production. However, academics can continue to address the integrity of the inputs by verifying their originality, content moderators can ensure that information is impartial and equitable distributed. So spotting illegal AI generated copies that look like the creative workers help protect intellectual property.

Taken together, these improvements increase the level of social trust put in digital interactions, intensify the accountability, and increase the ethicalness of a digital ecosystem.

Our research has societal benefits but also acknowledges the computational

processes which have an environmental impact. Machine learning model training and deployment on big datasets take a lot of energy resources which, in turn, spill over to greenhouse gas emissions. To address this, we emphasize the importance of adopting energy efficient practice such as using cloud platform powered by renewable energy source, as well as minimizing computation workflow.

Also, pre trained models and efficient algorithms are leveraged to save resources, and thus our work becomes more sustainable. We seek to do so by priming progress in AI research toward environmentally responsible methodologies while enabling advancement in technology, so that the work of scientists to advance AI technology does not lead to environmental impairment. Adding for the social impact and sustainability, this double emphasis calls for responsible innovation for current achievements in the field AI.

5.2.3 Ethical Aspects

We had ethical concerns at the core of every stage of our study. First, we put fairness first by resolving any possible biases in the dataset, and ensuring that the representation is balanced. Biased AI algorithms can yield discriminatory results that favor some while disadvantaging others, leading to a reinforcement of social injustices. We want to build inclusive and equitable models that cater to users from various demographics with a variety of data and rigorous assessment. Along the way, we also put great emphasis on accountability and transparency. We hope that our approaches are understandable, so that users can see how the model's forecasts make sense. If it's in reporting, education or public speaking, this is key to creating trust. Furthermore, in following through with data protection laws, no private information was commented in any manner or allowed to be used for unethical reasons. By staying true to those norms we have demonstrated that we can conduct moral AI practices that are applied constructively and protected from abuse, meaning this study is dedicated to the correct application of the results.

5.2.4 Sustainability Plan

Sustainability strategy: our research is sustainable towards its long-term sustainability and its responsible use of findings. One of the main priorities is to reduce the environmental effect of AI development. With our use of the cloud platforms powered by renewable energy sources and energy efficient technology, we

are able to lessen the carbon footprint of our research. It also shows that a healthy development of these models needs to balance environmental stewardship and innovation by improving model designs to reduce computing costs without loss of performance.

In addition, the ideas surrounding cooperation and open access are emphasized in our sustainability plan. It sends the message to community about research results, data and code, and promotes more creativity and helps avoiding unnecessary repetition of work. However, the use of such a strategy assures effective use of resources and accelerates development of the sector. This research encourages education and public awareness of moral practices of AI technologies that will maintain its beneficial effects across time through helping society adapt to technological breakthroughs in a responsible way.

5.3 Project Management and Financial Analysis

- Planning the Thesis and Writing the Proposal: The research began with a phase of detailed planning, during which some specific research objectives, research methods, evaluation parameters and a timeline were defined. This is a benefit in that the thesis is directed toward clear, measurable objectives that can be tracked through the thesis. This upload also contained thesis proposal, with research questions, theoretical framework, background literature, and possible give back to academic field.
- Supervision and guidance: During the long developing of this thesis project we were helped by our supervisor who supervised the idea and made sure that all goes correct as it should. It's good to have the supervisor to give us regular feedback to know what we are doing and what are supposed to do.
- Literature Review and Theoretical Foundation: In order to understand this context, further previous research and fill knowledge gaps, the study developed a literature review that informed the theoretical framework of the study. This extensive review informed the research design, hypotheses and the overall framing of analysis within a wider academic context.

- Data Collection and Analysis: Very rigid collection and understanding of data procedures were carefully planned and executed to ensure accuracy and reliability. We adhered to all ethical guidelines, maintained data integrity throughout preprocessing to final evaluation. This was an important step since it determined of the quality of the findings.
- Financial Analysis and Budgeting: To deal with potential financial need, a project budget was set based on software, data processing and contingency fund expenses. Minimizing spending and maintaining resource quality required effective financial management.

The estimated costs for the project are presented into Table 5.3.1 below.

TABLE 5.3.1: Estimated Cost the Project

SN	Components	Estimated Cost (BDT)
1	Internet Connection	1,000 - 1,500
2	Hardware	4,000 - 5,000
3	Software and Tools	1,000 - 1,200
4	Cloud Storage	1,000 - 2,000
5	Contingency	2,000 - 3,000
Total Estimated Cost		9,000 - 12,700

5.4 Complex Engineering Problem

5.4.1 Complex Problem Solving

In this section, provide a mapping with problem solving categories. For each mapping add subsections to put rationale (Use Table 5.4.1). For P1, you need to put another mapping with Knowledge profile and rational thereof.

Addressing of COs, Knowledge Profile (K), and Attainment of Complex

Engineering Problems (EP):

5.4.1: Addressing of COs, K and EP

SN	EP Definition	Attainment	CO	Justification (with Knowledge Profile)	Pages
1	EP1: Depth of Knowledge Required	Yes	CO1, CO2, CO3	- Requires knowledge of mathematics, statistics (K2). - Applied NLP model (K4). - Designed workflow systems (K5). - Conducted literature reviews (K8).	**Page no:** [6-21]
2	EP2: Range of Conflicting Requirements	Yes	CO2	- Balancing dataset exploration for accuracy and developing efficient models.	**Page no:** [16-20]
3	EP3: Depth of Analysis Required	Yes	CO2	- Conducted an in-depth analysis to develop accurate and reliable solutions while exploring multiple approaches.	**Page no:** [15-21]
4	EP4: Familiarity with Issues	Yes	CO3	- Researched and gained understanding of AI content, its type.	**Page no:** [6-14]

5	EP6: Extent of Stakeholders and Conflicting Requirements	No	CO3	- Engaged stakeholders, including educators and students, to gather feedback and refine data requirements.	
6	EP7: Interdependence	No	CO3	- Not applicable in the scope of this study.	
7	-	Yes	CO4	- Created budgets and cost estimates for project management and financial analysis.	**Page no:** [36-38]

5.5 Summary

In Chapter 5, the engineering standards, societal impacts and strategies to solve complicated problem are discussed in detail to meet the requirements of the project. Powerful software libraries like NumPy, Pandas, TensorFlow, and the like, along with the capabilities of the cloud platforms like Google Colab, Kaggle, can be used to simplify data processing and data visualization as well as build models. It was recommended that hardware specification in Intel i5 processor, 16 GB Ram, GTX 1650ti GPU will meet the computational demands efficiently. The project itself is showcased as a technology that mitigates critical societal problems, such as fighting misinformation, preventing unethical AI practices as well as protecting intellectual property. At the same time, environmental concerns were a priority, including energy efficient computation and platform utilizing renewable powered cloud platforms that minimize carbon footprints.From a management point of view the project demanded in fine details of planning, ethical collection of data and rigorous financial analysis, with costs estimated between 9,000 to 12,700 BDT. A multidisciplinary approach combining expertise in mathematics, AI, and workflow systems was used to solve complex engineering problems relevant to balancing accuracy with efficiency and accommodating divergent stakeholder needs. Overall, the chapter shows that the project champions ethical, sustainable and impactful solutions.

Chapter 6

Conclusion

In this research we developed an efficient and accurate NLP model using LSTM architecture to distinguish between machine generated and human written content with 98.31% accuracy while cost of computations is negligible. For future work, the goal is to increase model generalizability with multilingual datasets, optimize real-time deployment, and widen the model to identify output from different AI systems.

6.1 Summary

For this research I tried to develop a strong natural language processing (NLP) model to distinguish between machine generated and piece of work written by human. To handle the issue of the authenticity of digital content and, more generally, the prevalence of AI generated material in our ever more AI empowered world, we use a Text Vectorization layer, an embedding layer, and an LSTM based architecture to create our model. We aimed to have a model that was both computationally efficient and accurate, so that it could be used in practice for content moderation, verification of media authenticity, and academic integrity among others. The experiments were impressive. The test accuracy achieved was 98.31% with precision at 98%, recall 99%, and an F1 score of 98%. We saw steady improvement during training, and the small difference between training and validation metrics confirmed the model trained well on generialising well. Another thing is that the confusion matrix allowed assessing the model's reliability to accurately determine between machine generated and human written text. Additionally, this study included ethical and sustainable practices. In doing so, we aimed for as little computational overhead as possible and kept our sights on wider societal impacts of machine generated content detection. We note some limitations of our research, including a simplified dataset and a focus on binary classification.

6.2 Limitation

Our main goal having been achieved to produce a reliable and efficient system for distinguishing content generated from machines from that written by humans, the research worked successfully. Through the use of advanced NLP techniques we helped advance the work of understanding how AI driven systems are being monitored and even regulated to maintain a transparent and trustful state in digital spaces. We found that our LSTM based model is effective, while maintaining high accuracy with reasonable computational complexity.

With only 1.65 million parameters, the model was capable of real world use, tradeoffs between performance and resource efficiency. It is ideal for applications such as education, journalism and digital content moderation because of this. Additionally, we emphasized that while not unique to text processing, the use of easy to understand metrics to evaluate model performance is crucial for technical and non technical stakeholders alike as we ensured its reliability.

Specifically, we also noted challenges in addressed dataset biases, scalability, and generalizability. We found the model to work well on our dataset, but further research is needed to understand how it performs with other datasets and multilingual datasets. We also viewed this a limitation of the binary classification approach, as it was not capable of separating out the outputs from different AI systems.

6.3 Future Work

This research opens the door for multiple routes for future exploration:

- ESG in High Scarcity Problems: One can take this work further on transformer based architectures e.g. LSTM to extract richer context, and reason the model better. This would improve generalizability over languages and structures as well as adding depth and increase the trade offs for the model.
- Dataset Expansion and Diversity: Future research will consider multilingual and cross domain datasets. The model would be more adaptable and would reduce biases, leading the model to perform better on more applications than

can be accommodated by the model.

- Real-Time Deployment: Exploiting quantization and pruning type techniques to reduce the size of the model so that it is suitable for real time usage on edge devices or cloud platforms are to be explored. It would make it more usable and useful in every day applications.
- Exploring Class-Specific Analysis: If we were to extend the model further to encompass other outputs than just decision from one AI system to another, then we could actually distinguish between the outputs from different AI systems. The accuracy in model will be higher and the text written via AI will generate more insight.

Moving forward, future research can make use of our work to develop AI generated content detection that is ethically, reliably, and sustainably, as digital world continues to change.

References

[1] Marimuthu, M., Abinaya, M., & Hariesh, K. S. (2018). A review on heart disease prediction using machine learning and data analytics approach. International Journal of Computer Applications, 181(18), 20-25. https://doi.org/10.5120/ijca2018918012

[2] Motarwar, P., Duraphe, A., & Suganya, G. (2020). Cognitive approach for heart disease prediction using machine learning. 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE). https://doi.org/10.1109/ic-ETITE47903.2020.194

[3] Turin, T. C., Shahana, N., & Wangchuk, L. Z. (2013). Burden of cardio-and cerebro-vascular diseases and the conventional risk factors in South Asian population. Global Heart, 8(2), 121-130. https://doi.org/10.1016/j.gheart.2013.03.002

[4] Yusuf, S., Rangarajan, S., & Teo, K. (2014). Cardiovascular risk and events in 17 low-, middle-, and high-income countries. New England Journal of Medicine, 371(9), 818-827.

[5] Yang, L., Wu, H., & Jin, X. (2020). Study of cardiovascular disease prediction model based on random forest in eastern China. Scientific Reports, 10(1), 5245. https://doi.org/10.1038/s41598-020-62298-6

[6] Houston, M. (2018). The role of noninvasive cardiovascular testing, applied clinical nutrition, and nutritional supplements in the prevention and treatment of coronary heart disease. Therapeutic Advances in Cardiovascular Disease, 12(3), 85-108. https://doi.org/10.1177/1753944718763525

[7] Alaa, A. M., Bolton, T., & Di Angelantonio, E. (2019). Cardiovascular disease risk prediction using automated machine learning: A prospective study of 423,604 UK Biobank participants. PLOS ONE, 14(5), e0213653. https://doi.org/10.1371/journal.pone.0213653

[8] Balakrishnan, M., Christopher, A. A., & Ramprakash, P. (2021). Prediction of cardiovascular disease using machine learning. Journal of Physics: Conference

Series, 1767(1), 012013. https://doi.org/10.1088/1742-6596/1767/1/012013

[9] Y. Ma et al., "AI vs. Human – Differentiation Analysis of Scientific Content Generation," 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:256826708

[10] K. Schaaff, T. Schlippe, and L. Mindner, "Classification of Human-and AI-Generated Texts for English, French, German, and Spanish." [Online]. Available: https://copyleaks.com/ai-content-detector

[11] P. Lertvittayakumjorn and F. Toni, "Human-grounded Evaluations of Explanation Methods for Text Classification," in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Stroudsburg, PA, USA: Association for Computational Linguistics, 2019, pp. 5194–5204. doi: 10.18653/v1/D19-1523.

[12] J. Ji et al., "Detecting Machine-Generated Texts: Not Just 'AI vs Humans' and Explainability is Complicated," Jun. 2024, [Online]. Available: http://arxiv.org/abs/2406.18259

[13] S. Wiegreffe, J. Hessel, S. Swayamdipta, M. Riedl, and Y. Choi, "Reframing Human-AI Collaboration for Generating Free-Text Explanations," in Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Stroudsburg, PA, USA: Association for Computational Linguistics, 2022, pp. 632–658. doi: 10.18653/v1/2022.naacl-main.47.

[14] T. Berber Sardinha, "AI-generated vs human-authored texts: A multidimensional comparison," Appl. Corpus Linguist., vol. 4, no. 1, p. 100083, Apr. 2024, doi: 10.1016/j.acorp.2023.100083.

[15] N. N. I. Prova, "Detecting AI Generated Text Based on NLP and Machine Learning Approaches," ArXiv, vol. abs/2404.10032, 2024, [Online]. Available: https://api.semanticscholar.org/CorpusID:269157398

[16] A. M. Elkhatat, K. Elsaid, and S. H. Almeer, "Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text," Int. J. Educ. Integr., vol. 19, pp. 1–16, 2023, [Online]. Available:

https://api.semanticscholar.org/CorpusID:261398391

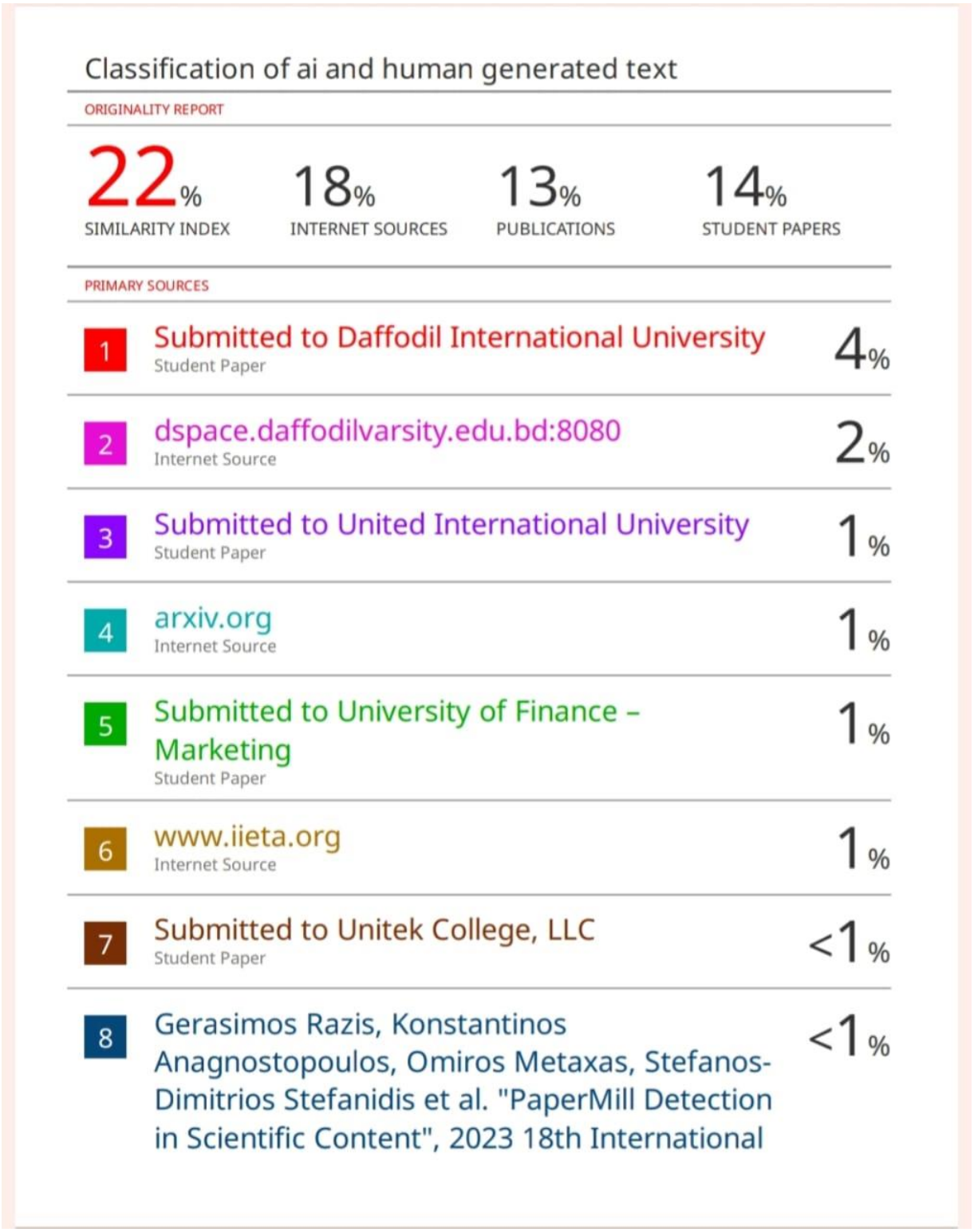
[17] N. Köbis and L. D. Mossink, "Artificial intelligence versus Maya Angelou: Experimental evidence that people cannot differentiate AI-generated from human-written poetry," Comput. Human Behav., vol. 114, p. 106553, Jan. 2021, doi: 10.1016/j.chb.2020.106553.

[18] E. Canhasi and R. Shijaku, "ChatGPT Generated Text Detection," doi: 10.13140/RG.2.2.21317.52960.

[19] G. P. Georgiou, "Differentiating between human-written and AI-generated texts using linguistic features automatically extracted from an online computational tool," ArXiv, vol. abs/2407.03646, 2024, [Online]. Available: https://api.semanticscholar.org/CorpusID:271038952

[20] G. Jawahar, M. Abdul-Mageed, and V. S. . L. Lakshmanan, "Automatic Detection of Machine Generated Text: A Critical Survey," in Proceedings of the 28th International Conference on Computational Linguistics, Stroudsburg, PA, USA: International Committee on Computational Linguistics, 2020, pp. 2296–2309. doi: 10.18653/v1/2020.coling-main.208.

[21] S. Herbold, A. Hautli-Janisz, U. Heuer, Z. Kikteva, and A. Trautsch, "A large-scale comparison of human-written versus ChatGPT-generated essays," Sci. Rep., vol. 13, no. 1, p. 18617, Oct. 2023, doi: 10.1038/s41598-023-45644-9.


Classification of ai and human generated text

ORIGINALITY REPORT

22%	18%	13%	14%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Daffodil International University Student Paper	4%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	2%
3	Submitted to United International University Student Paper	1%
4	arxiv.org Internet Source	1%
5	Submitted to University of Finance – Marketing Student Paper	1%
6	www.iieta.org Internet Source	1%
7	Submitted to Unitek College, LLC Student Paper	<1%
8	Gerasimos Razis, Konstantinos Anagnostopoulos, Omiros Metaxas, Stefanos-Dimitrios Stefanidis et al. "PaperMill Detection in Scientific Content", 2023 18th International	<1%