Truescript : Detection of AI-Generated and Human Written Texts

A Report Submitted in Partial Fulfilment of the Requirements for the

SN Bose Internship Program, 2024

Submitted by

Yashraj Poddar

Under the guidance of

Dr. Partha Pakray
Associate Professor
Department of Computer Science & Engineering
National Institute of Technology Silchar

Department of Computer Science & Engineering
NATIONAL INSTITUTE OF TECHNOLOGY SILCHAR
Assam

June-July, 2024

DECLARATION

TRUESCRIPT - DETECTION OF AI-GENERATED AND HUMAN WRITTEN TEXTS

I declare that the art on display is mostly comprised of my own ideas and work, expressed in my own words. Where other people's thoughts or words were used, I have properly cited and noted them in the reference materials. I have followed all academic honesty and integrity principles.

Yashraj Poddar

Yashraj Poddar

Department of Computer Science and Engineering
National Institute of Technology Silchar, Assam

ACKNOWLEDGEMENT

I would like to thank my supervisor, Dr. Partha Pakray, CSE, NIT Silchar, for his invaluable direction, encouragement, and assistance during this project. His helpful suggestions for this entire effort and cooperation are gratefully thanked, as they enabled me to conduct extensive investigation and learn about many new subjects.

My sincere gratitude to Mr. Advaitha Vetagiri and Mr. Yadagiri for their unwavering support and patience, without which this project would not have been completed properly and on schedule. This project would not have been feasible without them, from providing the data sets to answering any doubts I had to helping me with ideas whenever I was stuck.

Yashraj Poddar

Yashraj Poddar

Department of Computer Science and Engineering
National Institute of Technology Silchar, Assam

ABSTRACT

The digital world is evolving rapidly, and with it comes a new challenge: telling apart human writing from AI-generated text. As AI language models become more sophisticated, the line between human and machine-written content is blurring. This is where AI text detection steps in.

Our research presents the development and implementation of a web application designed to differentiate between AI-generated and human-written text. As AI language models become increasingly sophisticated, the ability to distinguish machine from human-authored content has become critical for maintaining authenticity across various domains, including academia, journalism, and online content moderation.

Our study employed a comparative analysis of several state-of-the-art natural language processing models for this classification task. We fine-tuned and evaluated the performance of BERT, RoBERTa, GPT-2, and a CNN-BiLSTM architecture. Each model was trained on a diverse dataset (HC3) of human-written and AI-generated texts. Performance was assessed using metrics such as accuracy, precision, recall, and F1-score.

The results indicated that BERT model exhibited superior performance in distinguishing between AI and human-authored text. This model was subsequently integrated into a web application, providing a practical tool for text origin detection.

This research contributes to ongoing efforts in AI detection and demonstrates the potential of machine learning in addressing challenges posed by advanced language models in content creation and verification. And even thought it is not perfect, the developed application offers a valuable resource for researchers, educators, and content moderators in an era of rapidly evolving AI-generated text capabilities.

Contents

Chapter 1

Problem Statement

Developing a Model to Distinguish Between AI-Generated and Human-Written Texts.

The primary objective of this project is to develop a robust machine learning model capable of distinguishing between AI-generated text and human-written text. The model will be trained on a diverse dataset containing examples of both types of text and will employ state-of-the-art natural language processing techniques to achieve high accuracy.

Chapter 2

Introduction

The rapid advancement of artificial intelligence (AI), particularly in natural language processing, has led to the increase in AI-generated text that is difficult to distinguish from human-written content. This development poses significant challenges across various sectors, including education, journalism, and online communication platforms. The lack of reliable methods to distinguish between AI-generated and human-authored text poses significant risks to the integrity of various forms of written communication, as highlighted in the paper ”Fighting Fire with Fire: Can ChatGPT Detect AI-generated Text?” [1] This deficiency compromises the authenticity of academic work, undermines the credibility of online information.

Research efforts such as ”The Science of Detecting LLM-Generated Texts [9],” ”RAIDAR: Generative AI Detection via Rewriting[7],” ”EAGLE: A Domain Generalization Framework for AI-generated Text Detection[2],” and ”Real or Fake Text? Investigating Human Ability to Detect Boundaries between Human-Written and Machine-Generated Text” highlight the need for robust AI text detection mechanisms. However, these studies exhibit several design flaws. ”The Science of Detecting LLM-Generated Texts” relies on overly complex methods that require significant computational resources and expertise. ”RAIDAR” and ”EAGLE” may face generalization issues, struggling to adapt to new, more sophisticated AI-generated texts.

Consequently, developing substantial AI text detection mechanisms becomes imperative to preserve the trustworthiness of human-generated content in an increasingly AI-capable world. By our studies, we find ways to detect if the text is AI-generated or human-written. We worked on researching this domain for ’Unmasking the Machine’ and achieving a ”Veritopia”—where truth and authenticity reign supreme, implying a grand vision of a society where digital communication is genuine.

Chapter 3

Dataset

The HC3 (Human ChatGPT Comparison Corpus) dataset consists of nearly 40K questions and their corresponding human/ChatGPT answers. The motivation for this dataset was to study ChatGPT's answers in contrast to human's answers. The questions range from a wide variety of domains, including open-domain, financial, medical, legal, and psychological areas. This wide-ranging approach means the dataset captures the diverse and complex nature of real-world writing. As a result, it's an excellent tool for creating AI detection systems that work well in many different situations and can handle various types of writing.

Following our work with the dataset, we compiled a comprehensive corpus comprising individual sentences and phrases, each accompanied by its corresponding classification label indicating human or machine authorship. We cleaned up and organized the data to get it ready for the next steps in our analysis. This dataset was really helpful because it gave us lots of different examples to work with. This was exactly what we needed to train our AI detection models and make them better at spotting the difference between human and AI-written text.

The dataset is originally linked as HC3 and can be found on Hugging Face.

Chapter 4

Methodology

4.1 Feature Augmentation

In this study, we utilized the HC3 dataset to classify text as either AI-generated or human-written. To enhance the dataset, we extracted six additional features: perplexity, burstiness, readability scores (Gunning Fog and Flesch-Kincaid), word density, and average line length. These features were derived after preprocessing the text column to ensure consistency and accuracy. Perplexity measures the predictability of a text, burstiness captures the variations in sentence structure, and readability scores indicate the ease of understanding the text. Word density and average line length provide insights into the textual characteristics that differentiate human-written content from AI-generated text.

4.2 Model Selection

We selected four models for this task: BERT[4], RoBERTa[6], GPT-2[8], and a CNN-BiLSTM[3] hybrid model. These models were chosen for their proven capabilities in natural language processing and text classification tasks. BERT and RoBERTa are transformer-based models known for their contextual understanding of language. GPT-2, a generative model, provides insights into the likelihood of a text being AI-generated. The CNN-BiLSTM model combines convolutional neural networks (CNN)[5] and bidirectional long short-term memory networks to capture both local and sequential patterns in the text. Each model was fine-tuned using the augmented dataset to improve their performance in distinguishing between AI-generated and human-written text.

4.3 Model Training

We conducted extensive experiments with different hyperparameters to identify the optimal settings for each model. The models were trained on the augmented dataset with the following labels: 0 for human-written text and 1 for AI-generated text. The best-performing model achieved an accuracy of 97.13%. The training process involved fine-tuning the pre-trained models on our dataset, adjusting learning rates, batch sizes, and other hyperparameters to maximize performance. The evaluation metrics included accuracy, precision, recall, and F1 score to ensure a comprehensive assessment of model performance.

To facilitate real-world application, we developed a graphical user interface (GUI) using React for the frontend and Flask for the backend. This interface allows users to input a single text and classify it as either AI-generated or human-written, leveraging the trained models for real-time prediction. The integration of the GUI with the classification models provides a user-friendly tool for detecting AI-generated text in various contexts.

Table 4.1: Model Performance Statistics

Model	**Accuracy**	**Precision**	**Test Loss**	**Recall**	**F1 Score**
BERT	0.9722	0.9735	0.0889	0.9706	0.97
CNN	0.9559	0.9520	0.0618	0.9517	0.9518
ROBERTA	0.9721	0.9629	0.1045	0.9818	0.97
GPT 2	0.9864	0.984	0.0460	0.984	0.99

Below are the plots representing the distribution of features with respect to the dataset we curated:

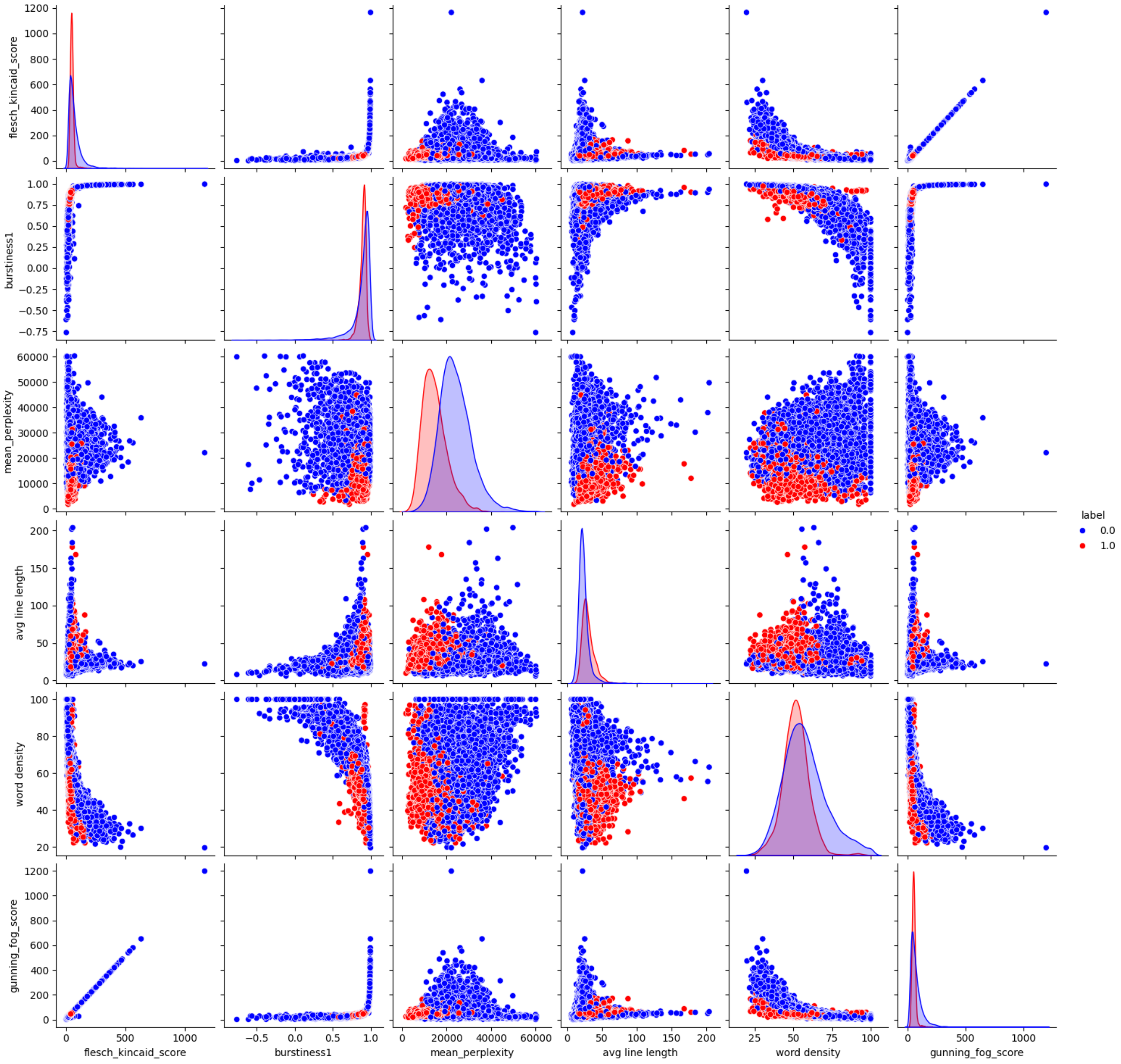


Figure 4.1: All Pair Plots for different features augmented

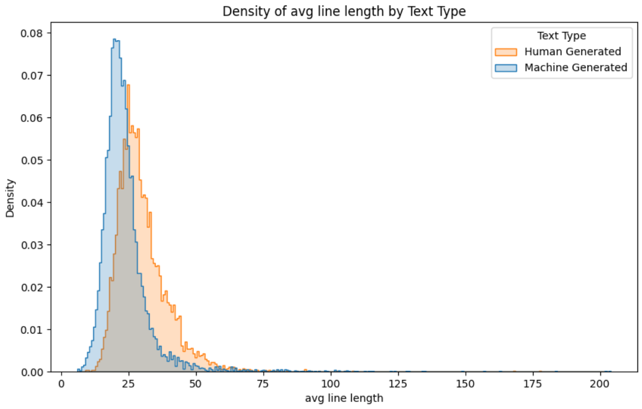


Figure 4.2: Density of Average Line by Text Type

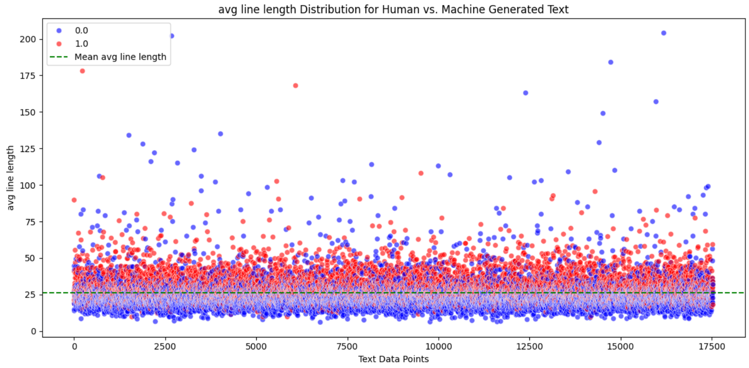


Figure 4.3: Average Line Length Distribution for Human vs Machine Generated Text

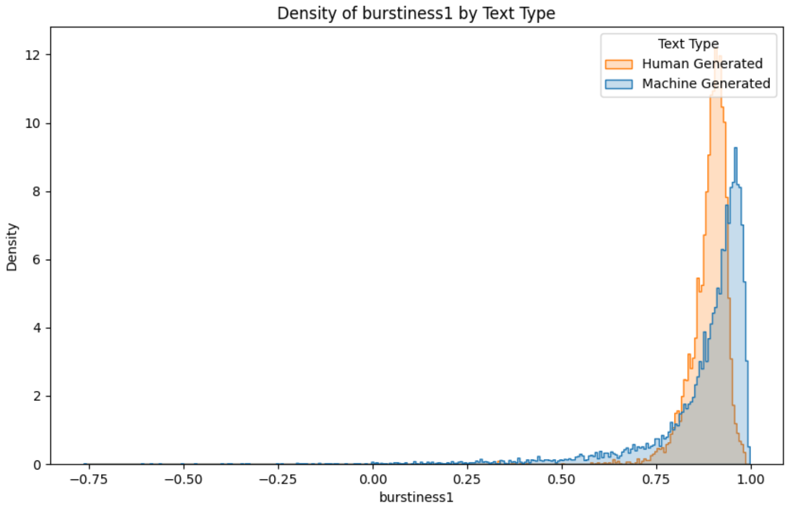


Figure 4.4: Density of burstiness by text type

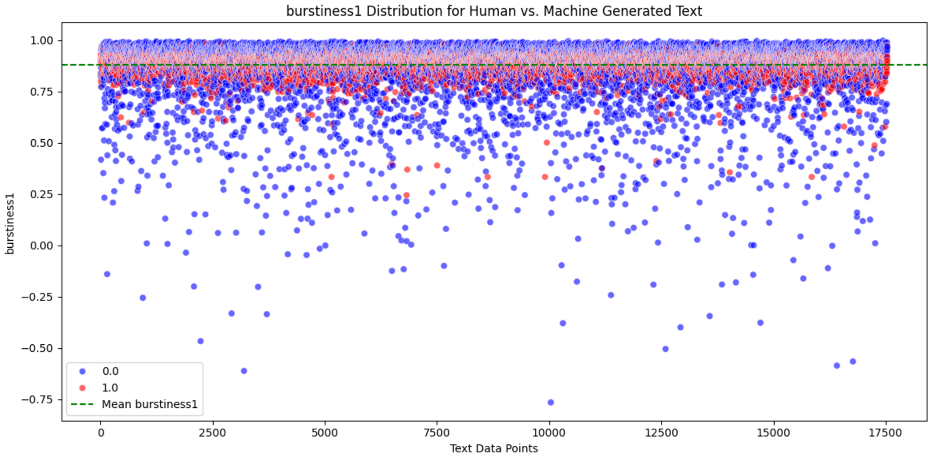


Figure 4.5: burstiness distribution for human vs machine generated text

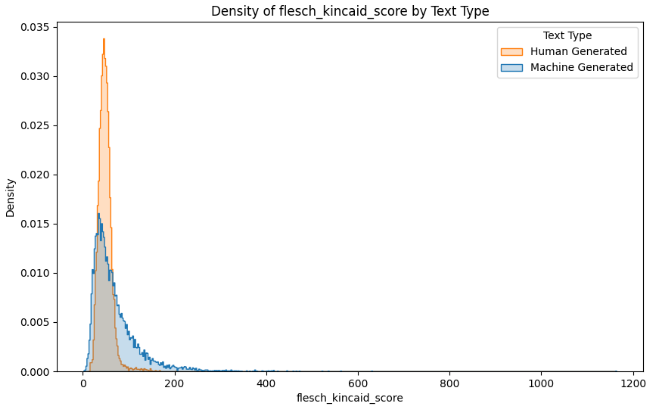


Figure 4.6: Density of flesch kincaid score by text type

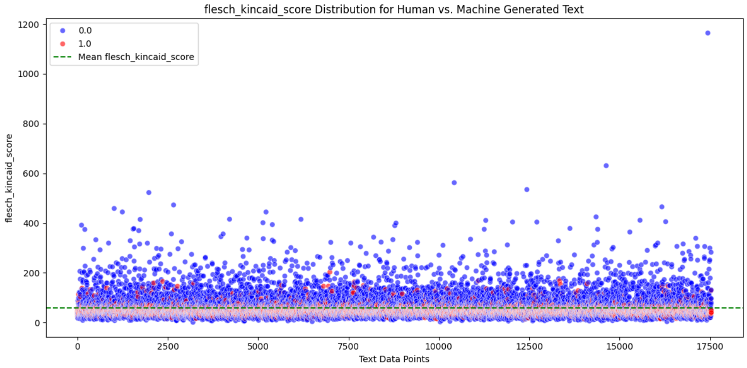


Figure 4.7: flesch kincaid score distribution for human vs machine generated text

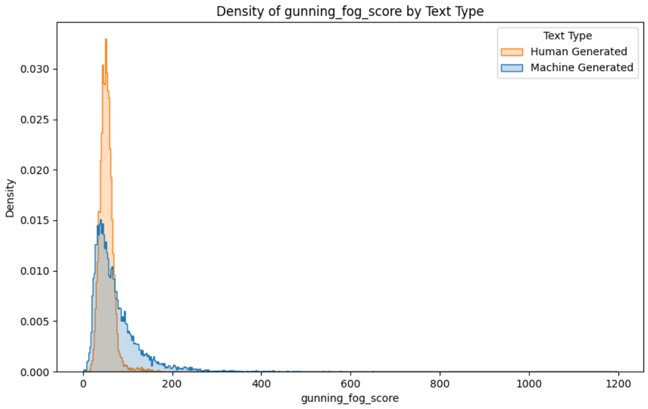


Figure 4.8: Density of gunning fog score by text type

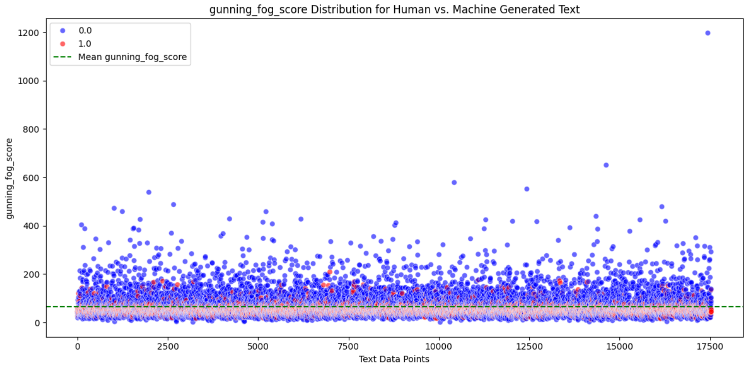


Figure 4.9: gunning fog score distribution for human vs machine generated tex

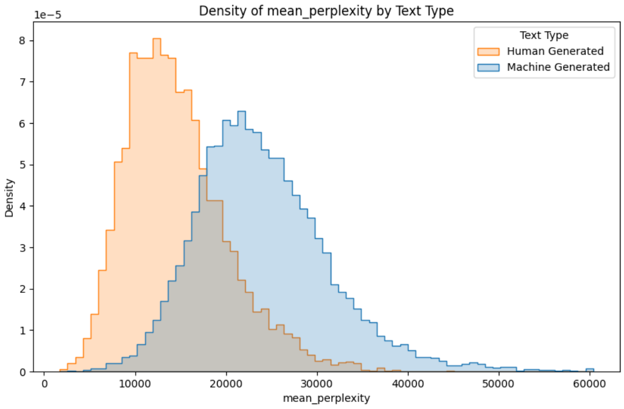


Figure 4.10: Density of mean perplexity by text type

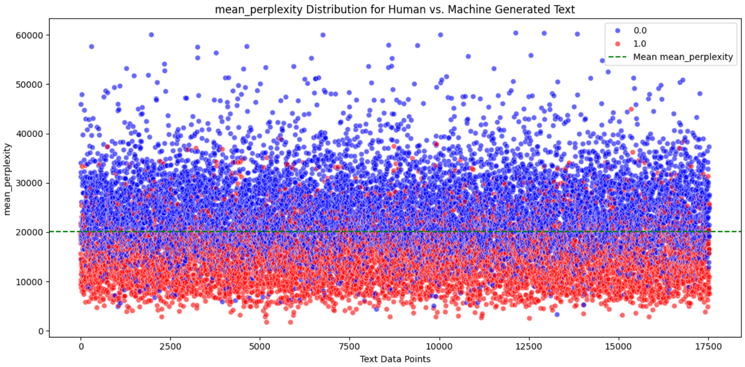


Figure 4.11: mean perplexity distribution for human vs machine generated text

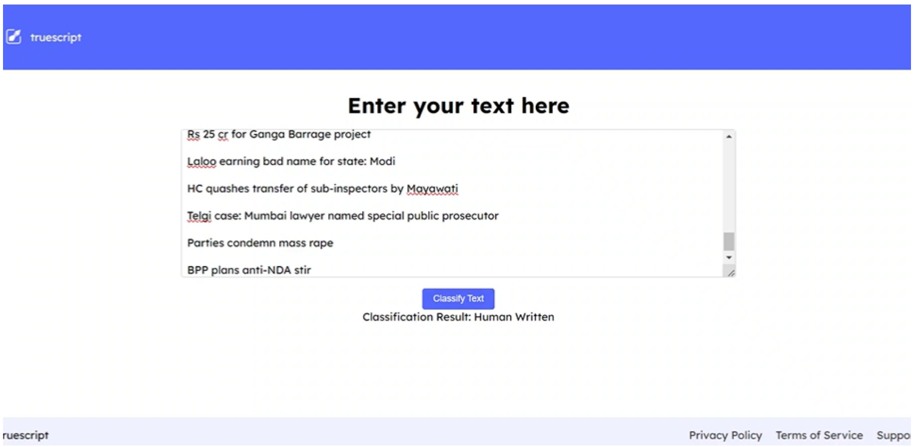
Chapter 5

Results

5.1 Trial 1

We fed the textbox with a human written article picked up from The Times of India, 2002 archive as our input text. The following shows the snapshot of the same.

Expected output: Human-written
Received output: Human-written



Enter your text here

Rs 25 cr for Ganga Barrage project

Laloo earning bad name for state: Modi

HC quashes transfer of sub-inspectors by Mayawati

Telgi case: Mumbai lawyer named special public prosecutor

Parties condemn mass rape

BPP plans anti-NDA stir

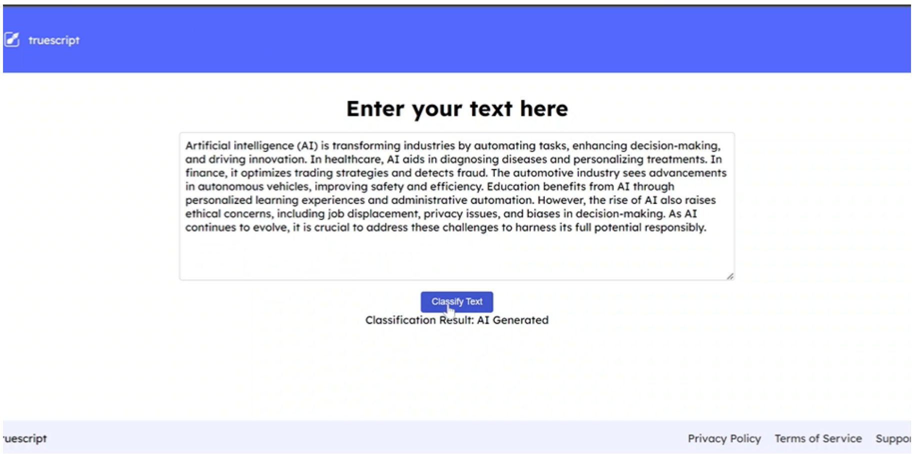
Classification Result: Human Written

Figure 5.1: Test of Human Written Text

5.2 Trial 2

We fed the textbox with a chatgpt generated paragraph as our input text. The following shows the snapshot of the same.

Expected output : AI Generated
Received Output : AI Generated



Enter your text here

Artificial intelligence (AI) is transforming industries by automating tasks, enhancing decision-making, and driving innovation. In healthcare, AI aids in diagnosing diseases and personalizing treatments. In finance, it optimizes trading strategies and detects fraud. The automotive industry sees advancements in autonomous vehicles, improving safety and efficiency. Education benefits from AI through personalized learning experiences and administrative automation. However, the rise of AI also raises ethical concerns, including job displacement, privacy issues, and biases in decision-making. As AI continues to evolve, it is crucial to address these challenges to harness its full potential responsibly.

Classification Result: AI Generated

Figure 5.2: Test of Human Written Text

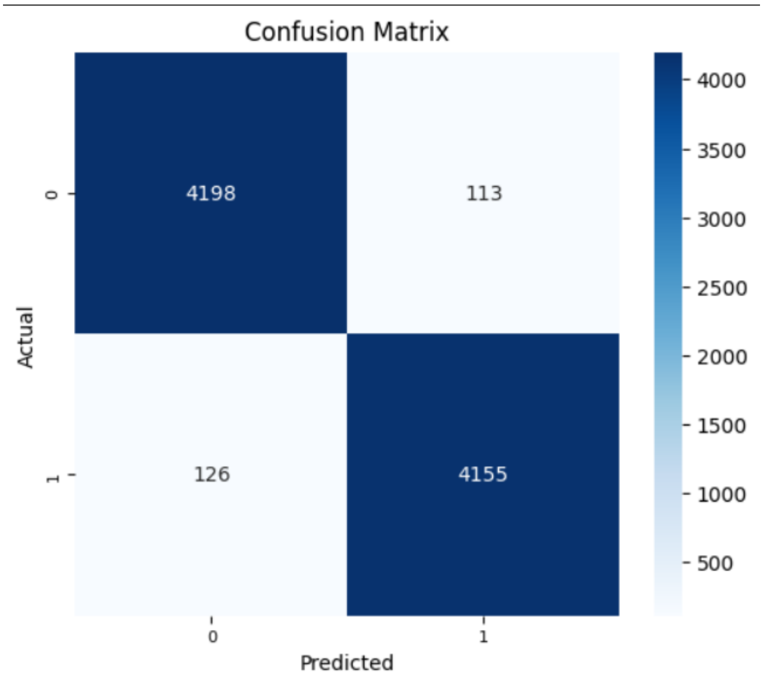


Figure 5.3: Confusion Matrix

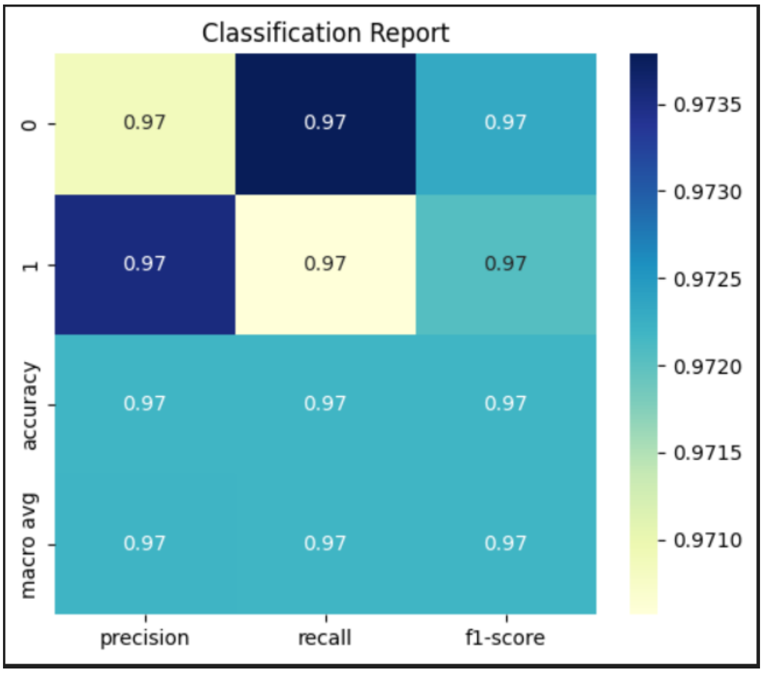


Figure 5.4: Classification Report

Chapter 6

Conclusion

In conclusion, our methodology for distinguishing AI-generated text from human-written content involved a comprehensive approach incorporating feature augmentation, model selection, and rigorous model training. By leveraging the HC3 dataset and enhancing it with six additional textual features, we were able to capture the nuanced characteristics differentiating AI and human text. The selection of advanced models such as BERT, RoBERTa, GPT-2, and a CNN-BiLSTM hybrid ensured robust performance, with fine-tuning tailored to our specific dataset. Through systematic experimentation with hyperparameters, we achieved a high accuracy rate of 97.13 %. The development of a user-friendly GUI integrating React and Flask further demonstrated the practical applicability of our models, offering real-time text classification capabilities. This multi-faceted approach not only underscores the efficacy of our methodology but also provides a scalable solution for the detection of AI-generated content in various contexts.

Chapter 7

Future Work

Looking ahead, there are several avenues to enhance our methodology for distinguishing AI-generated text from human-written content. First, incorporating additional features such as semantic similarity metrics and stylistic analysis can help capture deeper linguistic nuances, addressing the complexity observed in current methods. Additionally, exploring ensemble learning techniques by combining predictions from multiple models could improve classification accuracy and robustness, potentially overcoming issues of generalization.

Integrating advanced pre-trained language models like GPT-4 may offer significant performance gains by leveraging their sophisticated text generation capabilities. Another promising direction is the development of domain-specific classifiers tailored to different genres of text, such as news articles, social media posts, or academic papers. This approach could mitigate the high rates of false positives and negatives by focusing on context-specific characteristics.

Implementing a feedback mechanism in the user interface to allow users to contribute mislabeled or challenging examples could continuously refine and improve the model's accuracy over time, ensuring it stays up-to-date with evolving AI text generation techniques.

References

[1] A. Bhattacharjee and H. Liu. Fighting fire with fire: can chatgpt detect ai-generated text? *ACM SIGKDD Explorations Newsletter*, 25(2):14–21, 2024.

[2] A. Bhattacharjee, R. Moraffah, J. Garland, and H. Liu. Eagle: A domain generalization framework for ai-generated text detection. *arXiv preprint arXiv:2403.15690*, 2024.

[3] J. P. Chiu and E. Nichols. Named entity recognition with bidirectional lstm-cnns. *Transactions of the association for computational linguistics*, 4:357–370, 2016.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[5] K. Fukushima. Neocognitron. *Scholarpedia*, 2(1):1717, 2007.

[6] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[7] C. Mao, C. Vondrick, H. Wang, and J. Yang. Raidar: generative ai detection via rewriting. *arXiv preprint arXiv:2401.12970*, 2024.

[8] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[9] R. Tang, Y.-N. Chuang, and X. Hu. The science of detecting llm-generated text. *Communications of the ACM*, 67(4):50–59, 2024.