A Hybrid Approach for Discriminating Human vs AI Generated Text

Abde Manaaf Ghadiali
G29583342

Gehna Ahuja
G35741419

Sagar Sheth
G32921700

Venkatesh Shanmugam
G27887303

***Abstract*—As the prevalence of Large Language Models (LLMs) continues to rise, so too do concerns over their potential misuse, such as in AI-plagiarism scenarios. To address this issue, we propose a novel framework for discriminating between human-written and LLM-generated text. Our approach leverages an ensemble of machine learning and deep learning algorithms, distinguishing itself from existing methods. We utilize a unique dataset comprising text generated by LLM and human authors across various topics. After extensive experimentation and evaluation, we demonstrate the effectiveness of our framework in accurately discerning between human and LLM-generated text.**

Keywords— Large Language Models (LLMs), AI-plagiarism, Ensemble Learning, Deep Learning, Discrimination, Text Generation, Machine Learning, Natural Language Processing (NLP), Sentiment Analysis.

I. Introduction

After the advent of Large Language Models (LLM), we observed an enhanced efficiency of frameworks of various industries which adopted LLMs. However, we see another trend which talks about the misuse of LLMs. For example, students could use LLMs to generate essays to complete their assignments which is AI-Plagiarism and that could impact students' skill development which is a major concern of Educators these days. Therefore, our aim is to develop a framework that can distinguish between human-written and LLM-generated text.

The framework will be an ensemble of various Machine Learning and Deep Learning algorithms, rather than solely relying on a single model, and that is what sets us apart from the available approaches. The dataset will also be a unique combination of texts generated on diverse topics through LLM and human-written essays.

II. Literature Review

Accurately detecting LLM-generated text (LGT) is becoming increasingly critical as AI text generation models continue to evolve. Several studies emphasize the importance of high-quality and sufficient data for training effective LGT detection models. Research by [1] demonstrates that an extensive token length (word count) is crucial for accurate classification. [2], [3] and [4] further emphasize the need for diverse training data encompassing various writing styles and sources, such as social media, scientific publications, creative writing (stories, essays, poetries) and computer programs (Python scripts). This diversity helps models learn the nuances of human language across different contexts.

Deep learning approaches, Random-forest, Support Vector Machine (SVM), Long Short-Term Memory (LSTM) and particularly transformer architectures like RoBERTa, show promising results in LGT vs human-generated text (HGT) detection. Studies by [2] and [4] highlight the ability of these models to capture subtle language features and identify patterns within text data. This allows them to distinguish between the statistical regularities inherent in machine-generated text and the natural variations found in human-written language.

As AI text generation models become more sophisticated, LGT detection methods need to adapt. Research by [3] underlines the importance of generalizability, interpretability, and resilience against adversarial manipulation. Generalizable models can effectively detect LGT across various domains and writing styles. Interpretable models allow us to understand how the models make their decisions, which is crucial for building trust in their accuracy. Finally, robust detection methods should be resilient against adversarial techniques employed to evade detection, ensuring the integrity of the LGT detection system.

The study by [5] highlights the need for tailored LGT detection methods for specific domains, such as the medical field. Their research demonstrates that linguistic analysis, alongside machine learning models, is crucial for accurate detection in specialized domains where precise language usage is necessary. Medical documents often require a specific vocabulary and sentence structure that can differ from more general writing styles. By incorporating domain-specific knowledge, LGT detection methods can achieve greater accuracy in these specialized contexts.

By expanding on prior research, we aim to enhance the accuracy and effectiveness of methods which classify between HGT and LGT.

III. Methodology

Our methodology involved initial exploratory data analysis (EDA) on text generated by both AI and human sources. Following this, we fine-tuned Large Language Models (LLMs) and trained neural networks from the ground up. Finally, we employed an ensemble approach by selecting the top three models for further analysis.

A. Data Collection and Preprocessing

We sourced our dataset from [6], initially comprising extensive data. However, due to computational constraints, training on a dataset containing text from multiple LLMs became impractical. Therefore, we downsized the dataset to 52k records, evenly split between 26k from GPT-3.5 and 26k from human sources. Of these, 50k records were allocated for training, while 1k each were reserved for validation and testing purposes.

B. Exploratory Data Analysis

For conducting an analysis on both LGT and HGT, we explored various aspects, including the frequency of stop words used by the LLM and humans, examined the usage of punctuation marks and Sentiment analysis. This analysis aimed to provide insights into differentiating between AI-generated and human-written text, thereby validating or refuting certain hypotheses.

1) Stop Words Usage

We assessed the count of stop words in each record and plotted them on the x-axis, while the word count of each record was represented on the y-axis. Additionally, we utilized two distinct colors on the graph to distinguish between the two classes. This approach allowed us to investigate whether AI tends to use more stop words compared to humans.

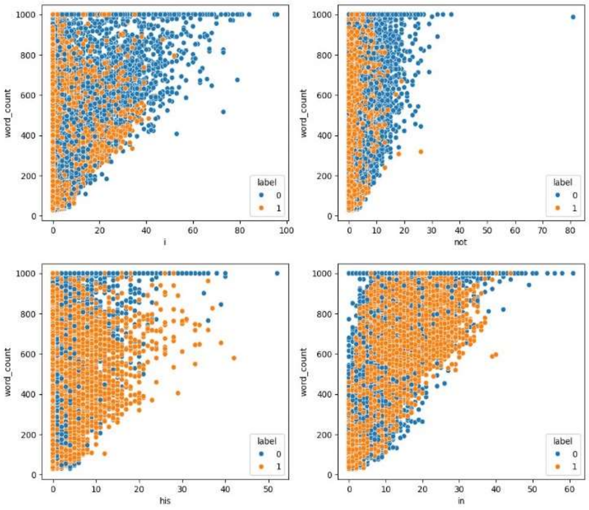


Figure 1: Scatter Plot of count of a few stop-words with respect to each records word count.

In the scatter plot, each point represents a record, with the x-axis indicating the word count of the record and the y-axis representing the count of selected stop-words. The labels distinguish between human-generated (Label 0) and LLM-generated (Label 1) text. For a comprehensive view of all stop words plotted, please refer to *Appendix - A*.

2) *Punctuation Marks*

Similarly, We evaluated the count of punctuation marks in each record and plotted them on the x-axis, while the word count of each record was depicted on the y-axis. Using two distinct colors for each of the classes.

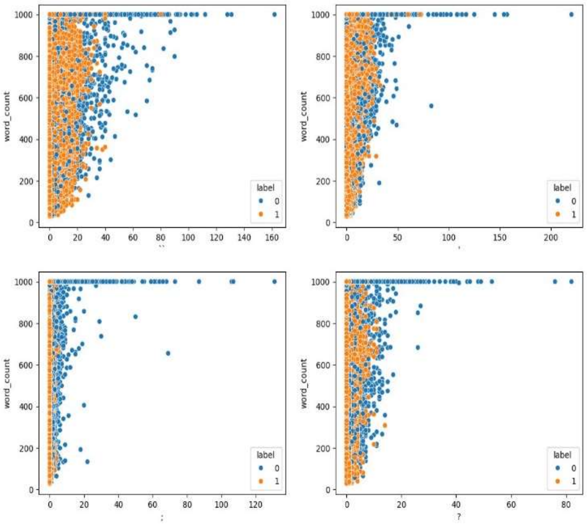


Figure 2: Scatter Plot of count of a few punctuation marks with respect to each records word count.

For a comprehensive view of all punctuation plotted, please refer to *Appendix - B*.

3) *Sentiment Analysis*

We conducted sentiment analysis on the dataset to discern the emotional tones exhibited by both humans and LLMs. We accomplished this task by employing the [7].

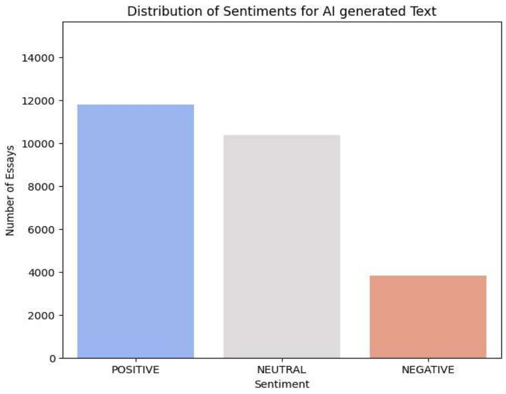


Figure 3: Distribution of Sentiments for LGT.

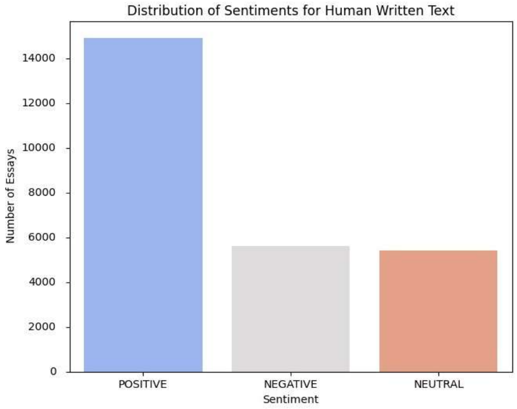


Figure 4: Distribution of Sentiments for HGT.

C. *Fine-Tuning Large Language Models (LLMs)*

We conducted fine-tuning on two LLMs, namely GPT-2 small and Roberta base, focusing solely on training the last transformer layer and ending dense layers. Due to computational constraints, we utilized a subset of 20k records for fine-tuning, while maintaining the validation and test datasets at the original size. Additionally, we tokenized sentences, ensuring that each record had a maximum of 512 sentences.

D. *Deep Neural Networks from Ground-up*

We developed custom Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) networks, and a basic feedforward Neural Network (NN) from scratch. Each of these models was trained on the entire dataset comprising 50k records, with a maximum word length of 1000 in each record.

E. *Ensembling Technique*

To achieve the most accurate results, we leveraged an ensemble approach. We trained several models, and then identified the top 3 performers based on their accuracy on both the test and validation datasets. We implemented a voting mechanism. Each model generated a prediction, and we tallied the votes for each class. The class with the most votes from our top models became our final prediction (*Appendix - C*).

For example, if first model predicts the Class '1' which means the data is generated from LLM, second model predicts class '0' which means data was written by Human and third model predicts class '1'. So, our final array for 1 record looks like [1,0,1], as we see class '1' has appeared twice so our final output is class '1'.

IV. Evaluation

As we trained multiple models, we assessed individual model's performance on the data. This evaluation process allowed us to make informed decisions regarding the selection of the top 3 models for ensemble.

A. Training Data Metrics

Model	Train Accuracy	Train Loss
GPT - 2 - Small	0.97	0.047
RoBERTa - Base	0.9928	0.0214
CNN	0.9954	0.016
RNN	0.5287	0.6916
LSTM	0.8788	0.2639
Deep Dense Model	0.528	0.6923

Model	Train ROC	Train F1
GPT - 2 - Small	0.99	0.9702
RoBERTa - Base	1	0.982
CNN	0.999	0.9969
RNN	0.52	0.1097
LSTM	0.98	0.2402
Deep Dense Model	0.52	0.00671

Class	Precision	Recall	F1 Score
0 - HGT	1	0.99	0.99
1 - LGT	0.99	1	0.99
accuracy			0.99
macro avg	0.99	0.99	0.99
weighted avg	0.99	0.99	0.99

B. Validation Data Metrics

Model	Val Accuracy	Val Loss
GPT - 2 - Small	0.958	0.114
RoBERTa - Base	0.976	0.07
CNN	0.954	0.1888
RNN	0.524	0.6922
LSTM	0.852	0.2905
Deep Dense Model	0.5278	0.6922

Class	Precision	Recall	F1 Score
0 - HGT	0.99	0.99	0.99
1 - LGT	0.99	0.99	0.99
accuracy			0.99
macro avg	0.99	0.99	0.99
weighted avg	0.99	0.99	0.99

C. Test Data Metrics

Model	Test Accuracy	Test Loss
GPT - 2 - Small	0.97	0.93
RoBERTa - Base	0.96	1.44
CNN	0.949	0.2
RNN	0.5217	0.6925
LSTM	0.8732	0.3007
Deep Dense Model	0.5257	0.6923

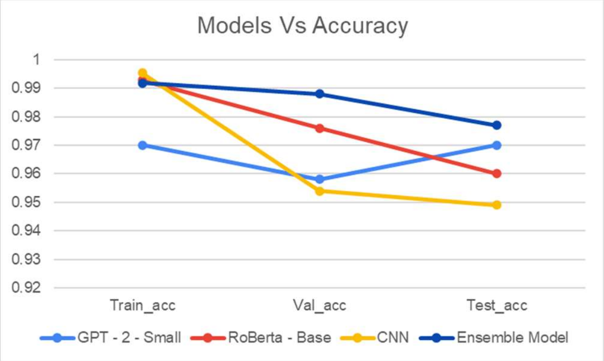
Model	Test Accuracy	Test Loss
GPT - 2 - Small	0.97	0.9739
RoBERTa - Base	1	0.9599
CNN	0.992	0.949
RNN	0.51	0.0951
LSTM	0.97	0.2303
Deep Dense Model	0.52	0.00199

Class	Precision	Recall	F1 Score
0 - HGT	0.98	0.98	0.98
1 - LGT	0.98	0.98	0.98
accuracy			0.98
macro avg	0.98	0.98	0.98
weighted avg	0.98	0.98	0.98

D. Ensemble Model Performance Analysis

Following an assessment of individual model performance, we selected the top 3 models for ensemble. Subsequently, we conducted a comparative analysis between the individual model performances and that of the ensemble model.

Model	Train Accuracy	Val Accuracy	Test Accuracy
GPT - 2 - Small	0.97	0.958	0.97
RoBERTa - Base	0.9928	0.976	0.96
CNN	0.9954	0.954	0.949
Ensemble Model	0.99176	0.988	0.977



The graph above clearly demonstrates that our ensemble model outperforms the individual models.

V. References

[1] G. Gritsay, A. Grabovoy and Y. Chekhovich, "Automatic Detection of Machine Generated Texts: Need More Tokens," 2022 Ivannikov Memorial Workshop (IVMEM), Moscow, Russian Federation, 2022, pp. 20-26, doi: 10.1109/IVMEM57067.2022.9983964.

[2] Gaggar, Raghav, Ashish Bhagchandani, and Harsh Oza. "Machine-Generated Text Detection using Deep Learning." arXiv preprint arXiv:2311.15425 (2023).

[3] Jawahar, Ganesh, Muhammad Abdul-Mageed, and Laks VS Lakshmanan. "Automatic detection of machine generated text: A critical survey." arXiv preprint arXiv:2011.01314 (2020).

[4] Kadhim Hayawi, Sakib Shahriar, Sujith Samuel Mathew," The Imitation Game: Detecting Human and AI-Generated Texts in the Era of ChatGPT and BARD",doi.org/10.1177/016555152412275.

[5] Wenxiong Liao, Zhengliang Liu, Haixing Dai, Shaochen Xu, Zihao Wu, Yiyang Zhang, Xiaoke Huang, Dajiang Zhu, Hongmin Cai, Tianming Liu, Xiang Li, "Differentiating ChatGPT-Generated and Human-Written Medical Texts: Quantitative Study", doi: 10.2196/48904

[6] https://www.kaggle.com/datasets/starblasters8/human-vs-llm-text-corpus

[7] https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment