Detecting AI-Generated Text with Pre-Trained Models using Linguistic Features

Annepaka Yadagiri, Lavanya Shree, Suraiya Parween, Anushka Raj, Shreya Maurya and **Partha Pakray**
Department of Computer Science & Engineering
National Institute of Technology Silchar, Assam, India, 788010
{annepaka22_rs,lavanya21_ug,suraiya21_ug,anushka21_ug,shreya21_ug, partha}@cse.nits.ac.in

Abstract

The advent of sophisticated large language models, such as ChatGPT and other AI-driven platforms, has led to the generation of text that closely mimics human writing, making it increasingly challenging to discern whether it is human-generated or AI-generated content. This poses significant challenges to content verification, academic integrity, and detecting misleading information. To address these issues, we developed a classification system to differentiate between human-written and AI-generated texts using a diverse *HC3-English dataset*. This dataset leveraged linguistic analysis and structural features, including part-of-speech tags, vocabulary size, word density, active and passive voice usage, and readability metrics such as *Flesch Reading Ease*, *perplexity*, and *burstiness*. We employed transformer-based and deep-learning models for the classification task, such as *CNN_BiLSTM*, *RNN*, *BERT*, *GPT-2*, and RoBERTa. Among these, the RoBERTa model demonstrated superior performance, achieving an outstanding accuracy of 99.73. These outcomes demonstrate how cutting-edge deep learning methods can maintain information integrity in the digital realm.

1 Introduction

The fast progress of Large Language Models (*LLMs*), like OpenAI's GPT-3 and GPT-4 (Sobieszek and Price, 2022) represent and other similar models developed by various organizations and start-up companies, has revolutionized the domain of Natural Language Processing (*NLP*) (Chowdhary and Chowdhary, 2020). These models, pre-trained on extensive text corpora, produce fluent and contextually appropriate writing. This aids in advancing several NLP activities, such as query translation, text categorization, and language translation. The ability to generalize without prior task-specific training is particularly noteworthy. The LLMs' adaptability in producing a range of writing styles, from academic to creative, without requiring domain-specific training is further shown by recent research by (Xu et al., 2021). Due to their versatility can be used in various contexts, including automated content creation, chatbots, and virtual assistants. However, this capability also brings significant challenges. With AI-generated text becoming increasingly prevalent, verifying the authenticity of content becomes increasingly complex, raising concerns over academic integrity and the spread of misinformation.

But moral dilemmas are associated with LLMs' sophisticated capabilities Bommasani et al., 2021. Their ability to write intelligible and contextually relevant content makes it easier for them to be abused, including spreading false information and fake news. These dangers damage society's views and undermine public confidence. Plagiarism, intellectual property theft, and the creation of false product evaluations are all issues that should worry businesses and customers alike (Radford et al., 2019). Additionally, LLMs can maliciously change web information, affecting political debate and public opinion. However, this capability also brings significant challenges. As AI-generated text becomes more prevalent, verifying the authenticity of content becomes increasingly complex, raising concerns over academic integrity and the spread of misinformation (Brown, 2020).

LLMs must be developed and implemented responsibly in light of these ethical issues. These approaches come with a complicated and diverse ethical environment. These problems must be addressed to realize the full benefits that appropriately deployed LLMs may bring to society. To do this, current research has shifted to developing detectors that can differentiate between text produced by computers and text written by people. These detectors act as a defense against possible LLM abuse.

Our research aims to develop a robust classifier

for differentiating human and AI-generated texts. We use the diverse HC3-English dataset, which contains several extracted features from both kinds of texts. We aim to analyze these features and train models to correctly identify the text's origin. Extracted features include all the POS tags, vocabulary size, word density, active and passive voice usage, Flesch Reading Ease score, Gunning Fog index, perplexity, and burstiness. We have taken more sophisticated models such as CNN_BiLSTM, RNN, BERT, and RoBERTa to evaluate their performance in this classification task. RoBERTa was the most accurate for this task, making this model the most effective for our task.

The key contributions of this paper are outlined as follows:

1. Feature Extraction Various linguistic analyses and statistical features were extracted from the dataset to enhance the identification of AI-generated text. These features include the Gunning Fog Index, perplexity, burstiness, readability scores (*such as Flesch-Kincaid*), *word density*, *average line length*, and *Pos tags*.

2. Model Training The extracted features were used to train DL models, including the RoBERTa model, demonstrating superior performance in discerning text generated by AI and human text.

The structure of this paper is as follows: Section 2 reviews existing research in the field of text classification, focusing on methods for distinguishing between human-written and AI-generated text. This section also presents various classification techniques and discusses previous SOTA model results, offering a comparative analysis of both types of texts. Section 3 outlines our approach to addressing the problem statement. It details the methods, including model selection, experimental setup, and implementation specifics. Section 5 discusses the performance of our models, highlighting the superior performance of the RoBERTa model. Section 7 summarizes the findings and concludes the paper. Section 6 suggests potential areas for further investigation and improvement based on the findings of this study.

2 Related Work

This chapter will review existing research on AI-generated text and explore the classification of texts generated by AI and humans.

2.1 AI-Generated Text Models

Recent advanced sophisticated LLMs, such as GPT-3, GPT-4, and other SOTA models like Pathways Language Model (*PaLM*), Gemini [1] by Google and Meta's Llama [2] (Touvron et al., 2023), have demonstrated significant capabilities in generating human-like text across various fields. These models are built upon NLP principles to generate coherent and contextually relevant responses based on user inputs such as language translation (Jiao et al., 2023), medicine, and education. These models have been effectively applied. Built on the Generative Pretrained Transformers (*GPT*) Language Model (*LM*), it is refined through human feedback and reinforcement learning, allowing them to understand user intent better and generate meaningful responses. Similarly, other LLMs like PaLM and Llama have pushed the boundaries of conversational AI, showcasing enhanced language understanding and generation capabilities. These models are trained on vast amounts of data to improve accuracy and safety in text generation. GPT-3, for example, was trained using *175 billion* parameters and *499 billion* tokens from crawled text data, positioning it as one of the most significant models at the time. In comparison, recent models like PaLM have utilized even larger datasets and parameters, with PaLM-2 incorporating *540 billion* parameters. While specific details regarding the training size of GPT-4 and other LLMs like Llama remain undisclosed, they follow similar large-scale training paradigms. Compared to other LMs, such as Bidirectional Encoder Representations from Transformers (*BERT*) (Yang et al., 2023), Robustly Optimized BERT Pretraining Approach (*RoBERTa*) (Liu et al., 2019). Text-to-text Transfer Transformer (*T5*) (Roberts et al., 2019), these recent models, including GPT-4, have shown significant improvements in understanding context and generating complex, human-like text, making them powerful tools for AI-generated text detection and content generation tasks.

2.2 Distinguishing Between Human- and AI-Generated Texts

The ability to distinguish between messages created by AI and humans becomes increasingly crucial

[1] https://gemini.google.com/app
[2] https://ai.meta.com/llama

as ChatGPT is utilized in more situations and its capabilities advance. Computers can already beat humans in detecting created texts as the quality of AI-made texts rises (Soni and Wade, 2023). Many tools are available to determine if a text has been developed by Artificial Intelligence (*AI*), such as GPTZero [3], AI Text Detection Tool [4], and GPT-2 Output Analyzer [5]. The foundation of these techniques is text pattern analysis. For example, one of the most widely used AI-detection programs, GPTZero, employs burstiness and perplexity to identify texts created by AI. These technologies still limit detection precision.

Approaches like XGBoost (Mindner et al., 2023), decision trees (Zaitsu and Jin, 2023), and transformer-based models (Mitrović et al., 2023; Guo et al., 2023) have been tested in recent research to identify texts created by AI. Developed a transformer-based classifier that could distinguish AI-generated text from human-generated language with an accuracy of 79%. Using decision trees that combined stylometric criteria specific to Japanese, including bigrams, punctuation placement, and frequency of function words, (Zaitsu and Jin, 2023) obtained 100% accuracy in recognizing Japanese texts. The qualities of AI-generated and human-generated responses to questions in English and Chinese were compared by (Guo et al., 2023). After optimizing a RoBERTa model for their texts, they obtained a 98.8 F1 score for the English responses. To solve the issue of detecting generated essays written in English (Shijaku and Canhasi, 2023) developed an XGBoost model that, when combined with a collection of manually created features and features produced by *TF-IDF*, obtained 98% accuracy. In their analysis of text summaries produced by humans and AI, (Soni and Wade, 2023) used DistilBERT [6] to reach 90% accuracy.

This research is the first to investigate an extensive set of features alongside state-of-the-art (*SOTA*) Transformer-based models, such as RoBERTa classifiers, to classify human and AI-generated text. Our results are compared against popular Machine Learning (*ML*) and Deep Learning (*DL*) models, including XGBoost, Random Forest (*RF*), and Multi-Layer Perceptron (*MLP*).

[3]https://gptzero.me/
[4]https://writer.com/ai-content-detector/
[5]https://openai-openai-detector.hf.space/
[6]https://huggingface.co/docs/transformers/model_doc/distilbert

3 Propsed Methodology

Problem statement

The sophistication with which AI-generated text proliferates has made it difficult to discern whether it is human or AI-generated text. This problem has real-world applications in domains where reliable text source classification is essential, such as spotting fraudulent emails and false news.

Given two text samples, the objective is to develop a binary classification model that can accurately identify which text is generated by AI and which is human-authored. This classification problem can be mathematically formalized as follows: Let x_i represent a text sample, where i denotes the index of the sample in the dataset. The objective is to categorize the text with a specific label y_i to each x_i, where:

$$y_i = \begin{cases} 0 & \text{if } x_i \text{ is Human-generated} \\ 1 & \text{if } x_i \text{ is AI-generated} \end{cases}$$

The problem then becomes one of estimating a function $f : X \rightarrow Y$, where X is the space of all possible text samples and $Y = \{0, 1\}$ is the set of possible labels.

Table 1: Comparison of Baseline Models and Proposed Models

Models	Training		Testing	
	Accuracy	F1 Score	Accuracy	F1 Score
Baseline Models				
CNN BiLSTM	75.0	75.0	76.34	76.35
RNN	68.0	68.24	67.64	67.66
BERT	67.0	67.24	64.92	64.34
GPT-2	83.0	83.0	82.88	82.90
RoBERTa	84.0	84.27	83.41	83.42
Proposed Models				
CNN BiLSTM	99.0	99.0	99.36	99.36
RNN	98.0	98.0	96.0	98.0
BERT	98.0	98.0	98.64	98.66
GPT-2	99.56	99.58	99.28	99.30
RoBERTa	**99.76**	**99.76**	**99.73**	**99.73**

3.1 Dataset Description

The HC3-English dataset comprises approximately 40,000 inquiries alongside corresponding responses from ChatGPT and humans. This dataset aims to facilitate a comparative analysis between ChatGPT's outputs and those generated by human respondents. The inquiries span multiple domains, such as open-domain discussions, finance, health, law, and psychology. Table 3 presents the meta-information regarding the HC3-English dataset utilized for training, with a breakdown as depicted in Table 2. This comprehensive scope enables

the dataset to encapsulate the intricate and varied essence of authentic writing. Consequently, it is an exceptional resource for developing AI detection systems that function effectively across diverse contexts and accommodate various writing styles.

Table 2: HC3-English Dataset Statistics

	Train Dataset Size(90%)	Test Dataset Size(10%)
Human:0	21890	2432
AI:1	21890	2432

3.2 Feature Extraction Techniques

3.2.1 Flesch Reading Ease

The Flesch Reading Ease score is a metric used to assess the readability of a text (Shijaku and Canhasi, 2023; Kincaid, 1975). It is determined by evaluating two main factors:

Average Sentence Length (*ASL*): The average sentence length in terms of word count.

Average Syllables per Word (*ASW*): The mean number of syllables found in each word. The Flesch Reading Ease score (Boudjella et al., 2017) is calculated using the following formula:

$$206.835 - (1.015 \times \text{ASL}) - (84.6 \times \text{ASW}) \quad (1)$$

Where: $\text{ASL} = \frac{\text{Total Words}}{\text{Total Sentences}}$, $\text{ASW} = \frac{\text{Total Syllables}}{\text{Total Words}}$

A higher score signifies more excellent readability, with scores ranging from 0 to 100, where higher scores are easier to read.

Sample Text: "*Salt is good for not dying in car crashes, and car crashes are worse for cars than salt. Some places use other things, but salt is really cheap compared to most alternatives, although sand is pretty good.*" Calculate the text. The Total Words: 34, Total Sentences: 3 Total Syllables (*estimated*): 48, ASL = $\frac{34}{3} \approx 11.33$ and ASW = $\frac{48}{34} \approx 1.41$ Flesch Reading Ease = $206.835 - (1.015 \times 11.33) - (84.6 \times 1.41) \approx 63.46$ Based on this score, the sample text has a moderate readability level, suitable for readers with some education. The average Flesch Reading Ease for the text type is shown in Figure 1, and the distribution of Flesch Reading Ease for human-generated vs AI-generated text is illustrated in Figure 2.

3.2.2 Gunning Fog Index

The Gunning Fog Index estimates the years of formal education required for a reader to understand a text on the first attempt (Kumarage et al., 2023). It considers the number of complex words in the text (*words with three or more syllables*). The formula is:

$$0.4 \times \left(\text{ASL} + \frac{\text{\% of Complex Words}}{100} \right) \quad (2)$$

Where: $\text{ASL} = \frac{\text{Total Words}}{\text{Total Sentences}}$, Percentage of Complex Words $= \left(\frac{\text{Number of Complex Words}}{\text{Total Words}} \right) \times 100$

The above Sample text has been taken. The Total Words: 28, Total Sentences: 2, Complex Words: 0, ASL = $\frac{28}{2} = 14$ and Percentage of Complex Words = $\frac{0}{28} \times 100 = 0\%$ then The Gunning Fog Index is = $0.4 \times (14 + 0) = 5.6$ A higher index indicates a more difficult text. The average of the Gunning fog index as shown in Figure 3 and the distribution of Gunning fog index for Human vs. AI-Generated Text is illustrated in Figure 4.

3.2.3 Perplexity

Perplexity is a measure used in LMs to gauge the accuracy with which a probability distribution or model predicts a given sample (Mindner et al., 2023). It measures uncertainty, with lower values indicating better predictive performance. Mathematically, for a given probability distribution P over a sequence $x_1, x_2, \ldots, x_n$, the perplexity $PP(P)$ is:

$$PP(P) = 2^{H(P)} = \exp\left(\frac{1}{N} \sum_{i=1}^{N} \left(-\log_2 \left(P(x_i) \right) \right) \right) \quad (3)$$

Where: $H(P)$ is the entropy of the distribution. $P(x_i)$ is the probability of the word x_i in the sequence. Perplexity is often computed on the test data using pre-trained LMs, like GPT-2. Lower perplexity implies that the model has greater confidence in its predictions.

The above Sample text has been taken. Assume we have an LM that assigns the following probabilities to the words in the sequence: $P(\text{"}Salt\text{"}) = 0.05$,$P(\text{"}is\text{"}) = 0.10$, $P(\text{"}good\text{"}) = 0.08$, $P(\text{"for"}) = 0.07$, ... (*remaining probabilities for other words in the sequence*)

The perplexity $PP(P)$ can be calculated by applying the probabilities to the formula. Note that the actual calculation would require the probabilities of all words in the sequence, but for simplicity, we use only a few probabilities in this example. The average of the perplexity of text type as shown in Figure 5, and distribution of perplexity for human vs. AI-generated text as illustrated in Figure 6.

Table 3: Meta-information of the HC3-English dataset.

HC3-English	# Questions	# Human Answers	# ChatGPT Answers	Source
All	24322	58546	26903	
reddit_eli5	17112	51336	16660	ELI5 dataset
open_qa	1187	1187	3561	WikiQA dataset
wiki_csai	842	842	842	Crawled Wikipedia (A.1)
medicine	1248	1248	1337	Medical Dialog dataset
finance	3933	3933	4503	FiQA dataset

3.2.4 Burstiness

Burstiness is a measure of the tendency of words to appear in clusters within a text (Mitrović et al., 2023). It quantifies how often certain words or terms are repeated in short intervals. One simple mathematical formulation of burstiness can be:

$$\text{Burstiness} = \sum_{i=1}^{V} \left(\frac{f_i^2}{T} \right) \quad (4)$$

Where V is the vocabulary size (number of unique words), f_i is the frequency of word i in the text, and T is the total number of words in the text. Burstiness measures the variability in word usage: texts with higher burstiness have words repeated more often close. This can indicate a repetitive or less diverse text structure. The average of burstiness scores by text type as shown in Figure 7, and distribution of burstiness for human vs AI-generated text is illustrated in Figure 8. The above Sample text has been taken. Assume we calculate the frequency of a few words: Frequency of "*salt*": $f_{\text{salt}} = 3$, Frequency of "*car*": $f_{\text{car}} = 3$, Frequency of "is": $f_{\text{is}} = 1$, V (*Vocabulary size*) = 17 (*assuming unique words in the text*) and T (*Total number of words in the text*) = 28. The burstiness can then be computed as:

$$\text{Burstiness} = \left(\frac{3^2}{28} \right) + \left(\frac{3^2}{28} \right) + \left(\frac{1^2}{28} \right) + \dots$$

In this case, words like "*salt*" and "*car*" might contribute more to the burstiness score due to their higher frequency. These metrics provide a way to quantify different aspects of the text, which can be helpful for both human analysis and as features in deep learning models.

The linguistic analysis within the lexical analysis focuses on surface-level features used to identify AI-generated text.

3.2.5 Average Line Length

The average line length represents the mean number of characters or words per line within a text dataset. The above sample text has been taken from the HC3-English dataset.

Average characters per line = (74 + 86) / 2 = 80 and *Average words per line* = (15 + 18) / 2 = 16.5

3.2.6 Vocabulary

Vocabulary denotes the collection of distinct words or tokens in a text dataset (Guo et al., 2023). The above sample text was taken from the dataset. Calculate the text's vocabulary, which consists of *31 unique words*.

3.2.7 Active Voice

Active voice refers to a sentence structure in which the subject carries out the action described by the verb. In the sample text taken from the dataset, there are 4 sentences, *3 of which are in active voice.*

$$\text{Average} = \frac{\text{No of active voice sentences}}{\text{Total number of sentences}} = \frac{3}{4} = 0.75 \quad (5)$$

3.2.8 Passive Voice

Passive voice is a sentence structure in which the subject receives the action carried out by the verb. The sample text taken from the dataset contains 4 sentences, *1 of which is in passive voice*. The formula for calculating the average passive voice is given below:

$$\text{Average} = \frac{\text{No of passive voice sentences}}{\text{Total number of sentences}} = \frac{1}{4} = 0.25 \quad (6)$$

3.2.9 Word Density

Word density quantifies the number of unique words present per text unit (Guo et al., 2023). It is calculated by multiplying the *vocabulary size by 100* and then dividing by the product of the number of lines and the average line length. The sample text provided was extracted from the dataset to calculate the total Vocabulary size: *31 unique words* and 2 lines in the text. Total characters: 87 + 117 = 204 characters, and Average line length: 204 / 2 = *102 characters.*

$$WD = \frac{100 \times \text{Vocabulary Size}}{\text{No of Lines} \times \text{Average Line Length}} = 15.2 \quad (7)$$

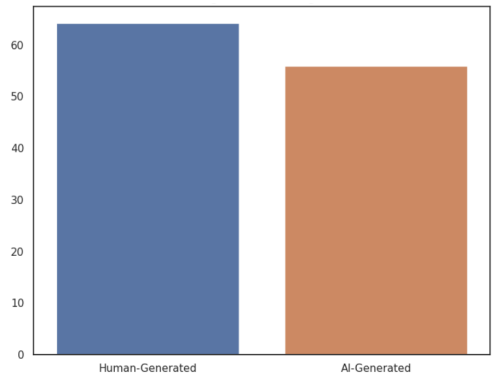

Figure 1: Average Flesch reading Ease

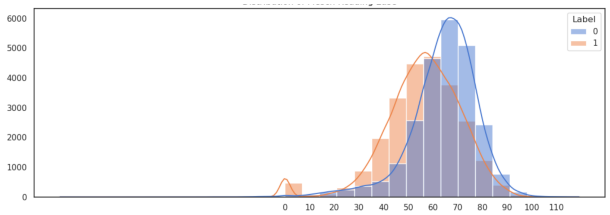

Figure 2: Distribution of flesh reading ease for Human vs. AI-Generated Text

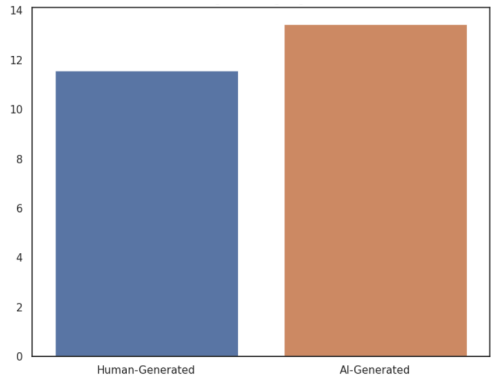

Figure 3: Average of Gunning fog index

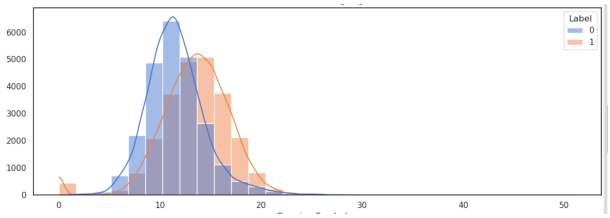

Figure 4: Distribution of Gunning fog index for Human vs. AI-Generated Text

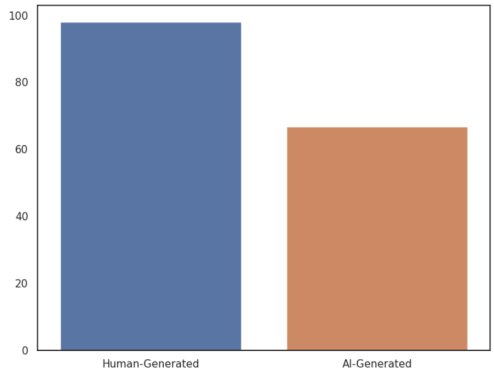

Figure 5: Average of perplexity of Text type

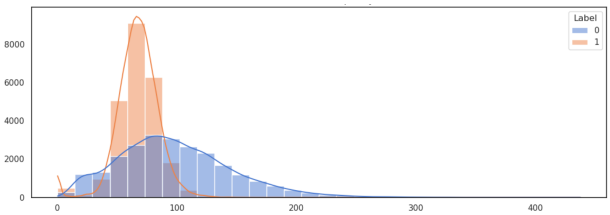

Figure 6: Distribution of perplexity for Human vs. AI-Generated Text

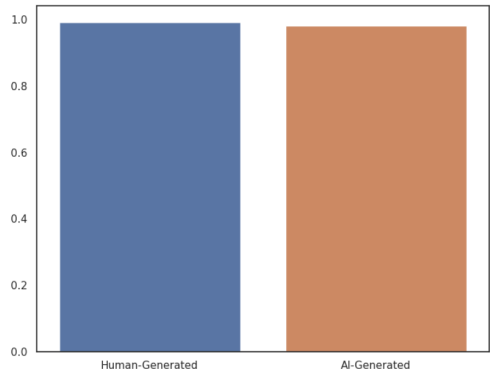

Figure 7: Average of Burstiness Scores by Text Type

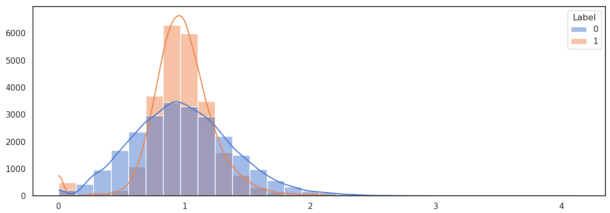

Figure 8: Distribution of Burstiness for Human vs. AI-Generated Text

3.2.10 POS Tags

Part-of-speech (*POS*) tags are labels assigned to each word in a text to denote its grammatical role, including categories such as '*NOUN*', '*VERB*', '*PUNCT*', '*DET*', '*PRON*', '*PROPN*', '*ADJ*', '*AUX*', '*ADV*', '*PART*', '*SCONJ*', '*NUM*', '*X*', '*INTJ*', '*ADP*', '*SYM*', '*SPACE*', and co-occurrence of conjunctions (*CCONJ*), which aids in understanding the syntactic structure and meaning of sentences. This study extracted POS tags from this dataset, incorporating *18 POS tag* features into the model training.

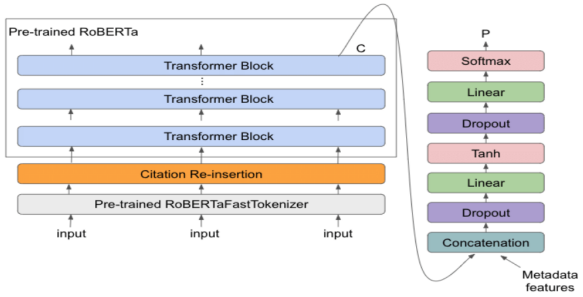

Figure 9: Proposed RoBERTa-Based Model Architecture

3.3 Model Selection

Recurrent Neural Networks (*RNNs*) (Elman, 1990) were initially leveraged for handling sequential data, making them suitable for tasks such as language modeling and time series prediction. Subsequently, a hybrid model, CNN_BiLSTM (Chiu and Nichols, 2016), was incorporated, wherein Convolutional Neural Networks (*CNNs*) and Bidirectional Long Short-Term Memory (*BiLSTM*) networks were combined. This approach allowed local features to be captured by CNNs while BiLSTMs were employed to model long-term dependencies in sequential data.

Moving to more advanced models, BERT (Devlin, 2018) Transformer-based model that reads text bi-directionally, understanding the context from both directions.

GPT-2 (Radford et al., 2019) Large-scale transformer-based language model designed to generate coherent and contextually relevant text. Then, Finally, RoBERTa (Liu et al., 2019) Optimized version of BERT, trained on more data with larger batches and longer sequences, and tokenized the data in a 12-layer model, such as RoBERTa-base. Hyperparameters, as shown in Table 4 tuning, Techniques such as grid search or random search were used to optimize the model performance.

The comparative analysis of these models provides valuable insights into the capabilities of different DL approaches for AI text detection.

Table 4: Hyperparameters Applied in Each Experiment

Parameter	**Value**
Activation Function	Sigmoid
Optimizer	AdamW
Loss Function	binary crossentropy
Learning Rate	$5 \times e^{-5}$
Batch Size	16
Number of Epochs	3
Learning Rate Scheduler	Linear
Dropout	0.2
ModelCheckpoint	Yes
EarlyStopping	Yes
Patience	3

4 Experimental Setup

The HC3-English dataset was partitioned into a training set of 90% and a test set of 10% comprising unseen data. The training was conducted over 3 epochs, fine-tuning the RoBERTa model using the hyperparameters outlined in Figure 4. The experiments were implemented in Python 3.10 and executed on a 3090 GPU with 64 GB of memory. To enhance the model's performance, linguistic features, and readability metrics were extracted and incorporated into the RoBERTa model. The architecture of the proposed model is illustrated in Figure 9. This model demonstrated superior performance, achieving an accuracy of 99.73. Metrics such as precision, recall, F1 score, and loss during the training process were recorded and are depicted in Figure 10. The model's accuracy improved significantly, starting from an initial 94.78 and reaching 99.73, demonstrating its effectiveness in detecting AI-generated text. Concurrently, the loss decreased from 0.25 to 0.02, indicating convergence toward accurate classification results. The training and test sets' average accuracy and loss values were computed. The training set recorded an average loss of 0.05, while the test set exhibited a slightly higher loss of 0.06. Regarding prediction accuracy, both the training and test sets achieved an average accuracy of 99.73, with a negligible decrease of 0.27 in the test set, showcasing the model's strong generalization capabilities.

5 Results and Discussions

Various DL and transformer-based models were utilized to evaluate training and testing accuracy, as demonstrated in the comparing baseline and proposed models as shown in Table 1. The RoBERTa

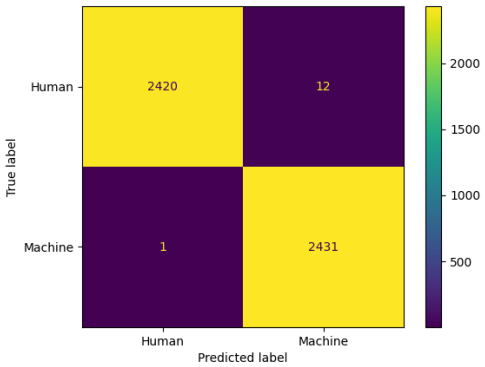

Figure 10: Confusion Matrix for the RoBERTa Model

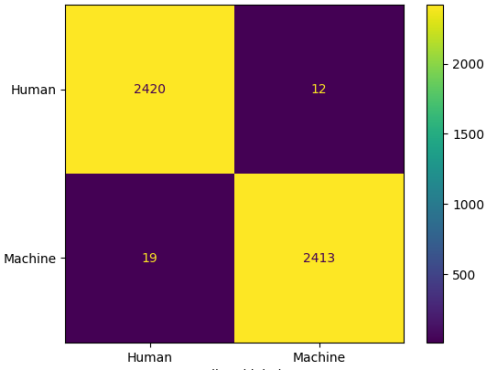

Figure 11: Confusion Matrix for the CNN_BiLSTM Model

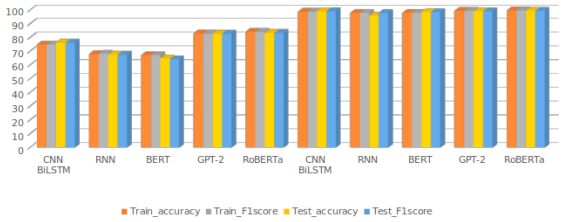

Figure 12: Results of Various Models: Training and Testing Accuracy and F1 Score Comparison

model achieved superior robustness and performance among all tested models, as illustrated in Figure 12. The confusion matrices, presented in Figures 10, 11 further highlight metrics such as precision, recall, and F1-score, culminating in an outstanding test accuracy of 99.73. This high accuracy indicates the model's exceptional effectiveness in distinguishing between human-written and AI-generated texts.

Several factors contributing to this success were identified. First, RoBERTa's advanced architecture was recognized as a significant contributor to its effectiveness. The model was pre-training on a vast corpus of diverse texts, enabling it to identify various linguistic patterns and styles. This comprehensive pre-training provided a solid foundation, allowing the model to excel across different text types.

6 Future work

In future research, we aim to gather additional data to expand the dataset to include a broader range of text types and sources. We also aim to explore other more effective feature engineering, investigating additional aspects that could optimize the model's architecture performance, such as syntactic and discourse features. Furthermore, we will develop more sophisticated models, experimenting with advanced DL techniques like graph neural networks and transformer-based models. Finally, we plan to deploy the model by developing a user-friendly interface to make the AI-generated text detection technology accessible to researchers, educators, and the general public. This step is crucial for improving the effectiveness and reliability of the technology across various application scenarios, enabling broader adoption and feedback-driven refinement.

7 Conclusion

In today's network security landscape, public opinion monitoring, and news media, identifying AI-generated text is growing progressively vital due to DL technologies' rapid advancements and extensive adoption. This study introduces an innovative AI text detection model built on the RoBERTa algorithm, providing a novel solution in this domain. Our approach integrates feature extraction, model development, and evaluation, demonstrating its efficacy with a test accuracy of 99.73. These results highlight the model's capability to differentiate between human-written and AI-generated text accurately, showcasing its potential to contribute significantly to NLP and LLM's responsible and ethical use. Through rigorous training and advanced DL techniques, the model exhibits high accuracy and low loss on the training set and stable and commendable performance on the test set. These findings are crucial for enhancing AI-generated text detection technology and serve as a

valuable resource for further research and practical applications in related fields.

Acknowledgements

The authors acknowledge the Centre for Natural Language Processing and the Department of Computer Science and Engineering, NIT Silchar, for their support and infrastructure.

References

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Aissa Boudjella, Mukti Sharma, and Deepti Sharma. 2017. Non-native english speaker readability metric: Reading speed and comprehension. *Journal of Applied Mathematics and Physics*, 5(6):1257–1268.

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Jason PC Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional lstm-cnns. *Transactions of the association for computational linguistics*, 4:357–370.

KR1442 Chowdhary and KR Chowdhary. 2020. Natural language processing. *Fundamentals of artificial intelligence*, pages 603–649.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jeffrey L Elman. 1990. Finding structure in time. *Cognitive science*, 14(2):179–211.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine. *arXiv preprint arXiv:2301.08745*.

JP Kincaid. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. *Chief of Naval Technical Training*.

Tharindu Kumarage, Joshua Garland, Amrita Bhattacharjee, Kirill Trapeznikov, Scott Ruston, and Huan Liu. 2023. Stylometric detection of ai-generated text in twitter timelines. *arXiv preprint arXiv:2303.03697*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lorenz Mindner, Tim Schlippe, and Kristina Schaaff. 2023. Classification of human-and ai-generated texts: Investigating features for chatgpt. In *International Conference on Artificial Intelligence in Education Technology*, pages 152–170. Springer.

Sandra Mitrović, Davide Andreoletti, and Omran Ayoub. 2023. Chatgpt or human? detect and explain. explaining decisions of machine learning model for detecting short chatgpt-generated text. *arXiv preprint arXiv:2301.13852*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Adam Roberts, Colin Raffel, Katherine Lee, Michael Matena, Noam Shazeer, Peter J Liu, Sharan Narang, Wei Li, and Yanqi Zhou. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *Google, Tech. Rep.*

Rexhep Shijaku and Ercan Canhasi. 2023. Chatgpt generated text detection. *Publisher: Unpublished*.

Adam Sobieszek and Tadeusz Price. 2022. Playing games with ais: the limits of gpt-3 and similar large language models. *Minds and Machines*, 32(2):341–364.

Mayank Soni and Vincent Wade. 2023. Comparing abstractive summaries generated by chatgpt to real summaries through blinded reviewers and text classification algorithms. *arXiv preprint arXiv:2303.17650*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Weijian Xu, Yifan Xu, Tyler Chang, and Zhuowen Tu. 2021. Co-scale conv-attentional image transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9981–9990.

Xiaofeng Yang, Fengmao Lv, Fayao Liu, and Guosheng Lin. 2023. Self-training vision language berts with a unified conditional model. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(8):3560–3569.

Wataru Zaitsu and Mingzhe Jin. 2023. Distinguishing chatgpt (-3.5,-4)-generated and human-written papers through japanese stylometric analysis. *PLoS One*, 18(8):e0288453.