Detecting AI-Generated Content: A Fine-Tuning Approach with Roberta-Base on HC3 Dataset

Xinhe Wu
University of Michigan, Ann Arbor
Email: xinhwu@umich.com

I. Introduction

A. Background and Motivation

The rapid development of large language models (LLMs) like ChatGPT has transformed natural language processing (NLP). As a conversational agent built on the GPT-3.5 architecture, ChatGPT delivers responses that are coherent, contextually accurate, and closely aligned with human intent. These capabilities have enabled applications across domains such as finance, medicine, and law.

Despite its advancements, ChatGPT raises societal concerns, including risks of misinformation and ethical misuse. The ability to distinguish between AI-generated and human-authored text has become critical, with applications in mitigating misinformation and ensuring AI transparency. Using the HC3 (Human ChatGPT Comparison Corpus) dataset [1], this study explores linguistic differences between human and ChatGPT responses and trains a RoBERTa-based binary classifier to detect AI-generated content.

B. Project Goal

The goal is to develop a system that accurately identifies human and AI-generated content. Specifically, we aim to:

- Analyze linguistic differences between human and ChatGPT responses.
- Train and evaluate a RoBERTa-based binary classifier.
- Provide insights to the detection of ChatGPT responses.

C. Literature Review

Recent studies highlight distinct characteristics of AI-generated text. The HC3 dataset reveals that ChatGPT responses are verbose, neutral in tone, and structurally rigid [1]. Herbold et al. [2] observed that ChatGPT essays, while structured and linguistically complex, lack the stylistic variety of human writing. Similarly, Mitrovic et al. [3] found AI-generated text to be polite, impersonal, and lacking emotional nuance, making it distinguishable from human content.

RoBERTa, introduced by Liu et al. [4], improves upon BERT by removing the next-sentence prediction objective and using dynamic masking, enabling it to capture long-range dependencies in text. RoBERTa has shown state-of-the-art performance on benchmarks like GLUE and has been successfully applied to tasks like fake news detection [5]. Its ability to capture subtle stylistic differences makes it an effective tool for detecting AI-generated content.

Transformer-based classifiers, including fine-tuned RoBERTa models, outperform traditional methods like perplexity-based approaches by identifying nuanced stylistic and linguistic features [3]. These strengths make RoBERTa an ideal candidate for this project.

II. Method

A. Problem Formulation

This study aims to address the problem of distinguishing human-generated content from AI-generated content using a supervised learning approach. The input to the model is a textual response (either human-generated or AI-generated), and the output is a binary label where `0` represents human-generated content and `1` represents AI-generated content.

The HC3 dataset is utilized, consisting of 24,322 data samples. Each sample includes a question and two corresponding answers: one authored by a human and the other generated by ChatGPT. The dataset is diverse, spanning multiple domains such as finance, medicine, and open-domain QA, ensuring robustness in cross-domain analysis. To prepare the dataset, the responses are tokenized and split into training (80%) and validation (20%) sets. The input text is truncated to a maximum of 128 tokens to ensure compatibility with transformer-based architectures.

The RoBERTa-base model is chosen for classification due to its robust text representation capabilities.

B. Methodology

The methodology consists of the following steps:

a) Exploratory Data Analysis (EDA):: EDA is performed to understand the characteristics of human and AI-generated responses. Metrics such as average answer length, vocabulary density, sentiment distribution, and part-of-speech (POS) tagging are analyzed. The results are visualized to uncover stylistic and linguistic patterns.

b) Data Preprocessing:: The dataset is first preprocessed to create structured input for the model. Each textual response is labeled and tokenized using the Hugging Face tokenizer for RoBERTa. To ensure consistency, padding and truncation are applied to limit the sequence length to 128 tokens. The processed data is then converted into a PyTorch-compatible format.

c) Model Training:: The RoBERTa-base model is initialized with a binary classification head for fine-tuning using the Hugging Face Trainer. The key hyperparameters for training include a learning rate of 1×10^{-5}, 3 epochs, a batch size of 4, and a weight decay rate of 0.01.

d) Evaluation:: The model's performance is evaluated using accuracy, precision, recall, and F1-score on the validation set. Additional metrics include the confusion matrix and classification report, which provide detailed insights into model performance.

III. Results

A. Exploratory Data Analysis (EDA)

The dataset analysis begins with visualizing key characteristics of human and ChatGPT-generated responses. This provides insights into the linguistic and stylistic patterns of each group.

a) Distribution of Sources and Average Answer Lengths:: Figure 1 shows the distribution of sources in the HC3 dataset alongside the average lengths of responses. The dataset is predominantly sourced from Reddit, while human answers exhibit greater variability in length compared to ChatGPT responses, which are more consistently distributed.

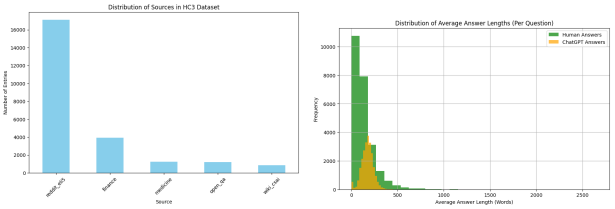

Fig. 1. (Left) Source distribution in the HC3 dataset. (Right) Distribution of average answer lengths for human and ChatGPT-generated responses.

b) Vocabulary Density and POS Tagging:: Figure 2 illustrates vocabulary density and part-of-speech (POS) tagging distributions. Human responses show higher variability in vocabulary density and greater diversity in POS tags, while ChatGPT responses are more consistent, emphasizing nouns and determiners.

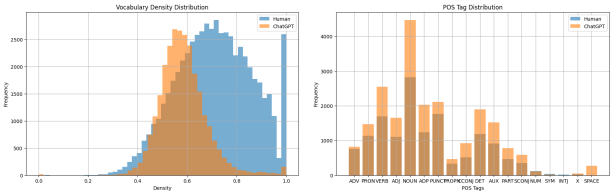

Fig. 2. (Left) Vocabulary density distribution. (Right) POS tag distribution for human and ChatGPT-generated responses.

c) Sentiment Analysis:: Figure 3 shows the sentiment distribution in human and ChatGPT-generated responses. Both groups exhibit predominantly negative sentiments, but ChatGPT demonstrates a slightly higher positivity rate, reflecting its training on polite and neutral text.

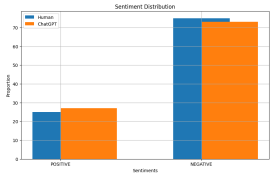

Fig. 3. Sentiment distribution for human and ChatGPT-generated responses.

B. Model Training and Evaluation Results

a) Training and Validation Loss:: Figures 4 display the training loss and validation metrics. The training loss decreases steadily, indicating effective optimization. Validation loss remains low, with validation accuracy consistently high across epochs.

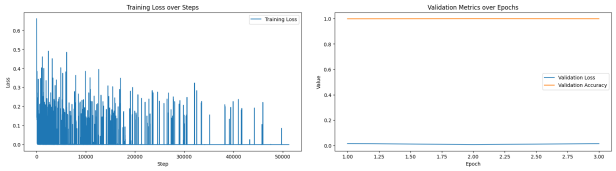

Fig. 4. (Left) Training loss over steps. (Right) Validation loss and accuracy over epochs.

b) Confusion Matrix and Classification Report:: Figure 5 presents the confusion matrix, confirming excellent performance in distinguishing human and ChatGPT responses. The classification report (Table I) demonstrates a precision, recall, and F1 score of 1.00 for both classes.

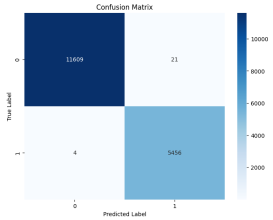

Fig. 5. Confusion matrix for classification results.

Class	Precision	Recall	F1-Score	Support
Human	1.00	1.00	1.00	11,630
ChatGPT	1.00	1.00	1.00	5,460
Accuracy: 1.00				

TABLE I
CLASSIFICATION PERFORMANCE METRICS.

C. Token Highlighting Analysis

a) Token Predictability Visualization:: Figures 6 compare the token predictability for human and ChatGPT-generated responses. Human text features more unpredictable tokens (red and purple), reflecting diverse linguistic richness. ChatGPT text is predominantly predictable (green and yellow), indicating a reliance on learned patterns.

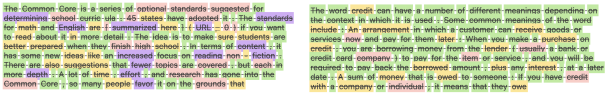

Fig. 6. (Left) Token predictability highlighting for human-generated text. (Right) Token predictability highlighting for ChatGPT-generated text.

IV. Conclusion

This study highlights the effectiveness of the RoBERTa-base model in distinguishing human and ChatGPT-generated text. Human responses exhibit greater linguistic diversity, while ChatGPT text is more structured and predictable. The fine-tuned model achieved near-perfect performance, with precision, recall, and F1-scores of 1.00. These findings demonstrate the potential of transformer-based models for detecting AI-generated content and provide insights into the stylistic differences between human and machine-generated text. Future work can explore newer AI models and enhance interpretability techniques.

References

[1] B. Guo *et al.*, "Human-chatgpt comparison corpus (hc3)," *arXiv preprint arXiv:2301.07597*, 2023.

[2] S. Herbold *et al.*, "Large-scale comparison of human- and ai-written essays," *Scientific Reports*, vol. 13, p. Article 45644, 2023.

[3] J. Mitrovic *et al.*, "Detecting short texts generated by chatgpt," *arXiv preprint arXiv:2301.13852*, 2023.

[4] Y. Liu *et al.*, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[5] K. Shu *et al.*, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD Explorations Newsletter*, 2020.