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 emergence of large language models (LLMs) like ChatGPT has revolutionized natural language processing (NLP), enabling applications across domains such as finance, medicine, and law. As a GPT-3.5-based conversational agent, ChatGPT generates responses that are coherent, contextually accurate, and human-like. However, its growing use raises societal concerns, including risks of misinformation and ethical misuse. Identifying AI-generated content has become essential for mitigating misinformation and ensuring transparency.

This study explores linguistic differences between human and ChatGPT responses using the HC3 (Human ChatGPT Comparison Corpus) dataset [1] and trains a RoBERTa-based binary classifier to detect AI-generated content.

B. Project Goal

This project aims to develop a system to accurately distinguish between human and AI-generated content by analyzing linguistic differences, training a RoBERTa-based classifier, and providing insights into detecting AI-generated text.

C. Literature Review

Research highlights distinct patterns in AI-generated text. ChatGPT responses tend to be verbose, neutral in tone, and structurally rigid [1], often lacking the stylistic variety and emotional nuance of human writing [2], [3]. RoBERTa, introduced by Liu et al. [4], enhances BERT by removing the next-sentence prediction task and employing dynamic masking, achieving state-of-the-art results in text classification tasks, including fake news detection [5].

Transformer-based classifiers like RoBERTa excel at capturing nuanced linguistic features, outperforming traditional methods such as perplexity-based approaches [3]. These strengths make RoBERTa a compelling choice for detecting AI-generated content in this project.

II. Method

A. Problem Formulation

This study addresses the challenge of distinguishing human-generated content from AI-generated content through supervised learning. The input is a textual response, and the output is a binary label: `0` for human-generated and `1` for AI-generated content.

The HC3 dataset, comprising 24,322 samples, includes questions with paired human and ChatGPT responses spanning diverse domains such as finance and medicine. The dataset is split into 80% training and 20% validation sets. Responses are tokenized and truncated to a maximum of 128 tokens for compatibility with transformer models. The RoBERTa-base model is selected for its robust text representation capabilities.

B. Methodology

The approach involves the following steps:

a) Exploratory Data Analysis (EDA):: Key characteristics of human and AI-generated responses, such as average answer length, vocabulary density, sentiment, and part-of-speech (POS) tagging, are analyzed and visualized to reveal linguistic patterns.

b) Data Preprocessing:: Textual responses are labeled and tokenized using the RoBERTa tokenizer. Padding and truncation are applied to limit sequences to 128 tokens, and the data is converted to a PyTorch-compatible format.

c) Model Training:: The RoBERTa-base model is fine-tuned with a binary classification head. Training is conducted using a learning rate of 1×10^{-5}, 3 epochs, a batch size of 4, and a weight decay rate of 0.01.

d) Evaluation:: Model performance is assessed through accuracy, precision, recall, F1-score, a confusion matrix, and a classification report, providing detailed insights into its effectiveness.

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.

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.

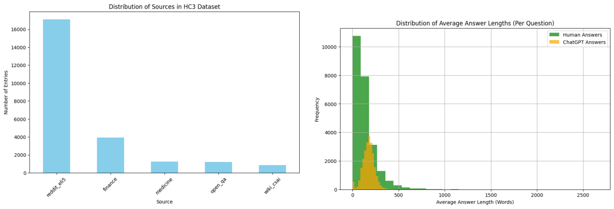


Fig. 1. (Left) Source distribution in the HC3 dataset. (Right) Distribution of average answer lengths for human and ChatGPT-generated responses.

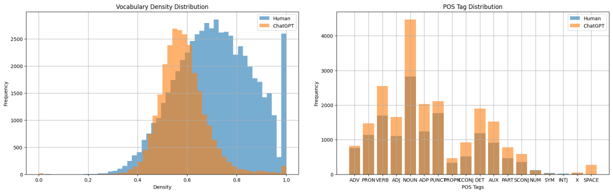


Fig. 2. (Left) Vocabulary density distribution. (Right) POS tag distribution for human and ChatGPT-generated responses.

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.

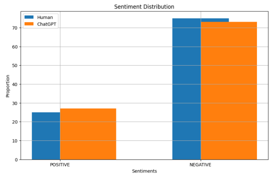


Fig. 3. Sentiment distribution for human and ChatGPT-generated responses.

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.

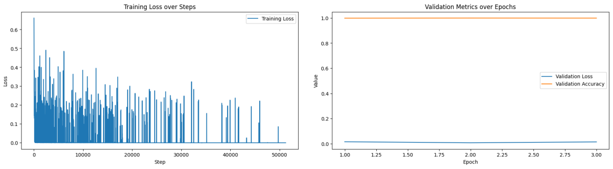


Fig. 4. (Left) Training loss over steps. (Right) Validation loss and accuracy over epochs.

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.

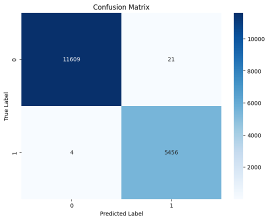


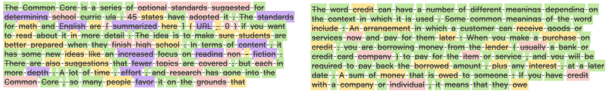
Fig. 5. Confusion matrix for classification results.

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.


The Common Core is a series of optional standards suggested for determining school curric ula . 45 states have adopted it . The standards for math and English are [summarized here] (URL _ 0) if you want to read about it in more detail . The idea is to make sure students are better prepared when they finish high school . In terms of content , it has some new ideas like an increased focus on reading non - fiction . There are also suggestions that fewer topics are covered , but each in more depth . A lot of time , effort , and research has gone into the Common Core , so many people favor it on the grounds that

The word credit can have a number of different meanings depending on the context in which it is used . Some common meanings of the word include : An arrangement in which a customer can receive goods or services now and pay for them later : When you make a purchase on credit , you are borrowing money from the lender (usually a bank or credit card company) to pay for the item or service , and you will be required to pay back the borrowed amount , plus any interest , at a later date . A sum of money that is owed to someone : If you have credit with a company or individual , it means that they owe

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.