AI vs Human Text Classifier using RoBERTa

Student: Akash Anand
Track: NLP / Transformers
Trimester: 3rd Trimester, Minor in Data Science and AI/ML
Submission: AI vs Human Writing Detection Capstone Project

Abstract

Nowadays AI tools are generating a lot of text and it is very hard to know if a text is written by a human or an AI. In this project I tried to solve this problem by training a model which can classify text into two types – human written or AI generated. For this I used RoBERTa transformer model and fine-tuned it on a dataset of human and AI text. I also tested another model called DistilBERT, but its performance was not as good as RoBERTa. The project followed a complete workflow: dataset collection and balancing, preprocessing with tokenizer, fine-tuning the transformer, evaluation using multiple metrics, and finally deployment using Gradio on Hugging Face Spaces. The deployed app allows anyone to check text in real time and see if it is human or AI with a confidence score.

Problem Statement

The availability of AI text generators like GPT-3 and GPT-4 is making it extremely easy for people to produce text automatically. While this has many advantages, it also comes with serious issues like:Students using AI for assignments and plagiarism.Fake information being spread without fact checking.Losing originality and human creativity.Difficulty in checking authenticity of research content.One of the most difficult parts of this project was to build a balanced dataset. If the dataset is not balanced (for example more human text than AI text), then the model will become biased and predict the majority class more often. This means the model will look like it is performing well, but in reality it is just biased. So preparing the dataset properly was one of the most important steps.
Another challenge was model selection. I tried DistilBERT because it is lightweight and faster, but the accuracy was much lower compared to RoBERTa. After comparing results, I decided to use RoBERTa-base as my final model since it gave more reliable classification.

Objectives

Build a machine learning model that can classify text into Human or AI-generated.
Fine-tune RoBERTa transformer model for this binary classification task.
Experiment with at least one other model (DistilBERT) for comparison.
Evaluate the model with metrics such as Accuracy, Precision, Recall, and F1-score.
Deploy the model using Gradio and Hugging Face Spaces so anyone can test it.Explore possible applications in education, research integrity, and online content verification.

Methodology

1. Dataset PreparationCollected text samples from both human-written sources (like essays, articles, blogs) and AI-generated content (produced by GPT models).Each entry was labeled as 0 (Human) or 1 (AI).Balancing was very important here, so I made sure the dataset contained almost equal amount of human and AI text.

2. **Preprocessing**

Used RoBERTa tokenizer to convert text into token IDs.Applied padding and truncation so that all sequences have same length (maximum 256 tokens).
Removed unnecessary symbols and cleaned short texts which had no meaning.

3. Model Selection and Training

I started with DistilBERT since it is smaller and faster. But the accuracy was not very good (it often misclassified short text).
Then I used RoBERTa-base, which is a stronger transformer model.Fine-tuned RoBERTa for sequence classification with 2 output labels (Human, AI).Training was done on Google Colab using Hugging Face Trainer API.
Optimizer: AdamW, with learning rate scheduling.Training was run for several epochs until validation accuracy stabilized.

4. **Evaluation**

I evaluated the model using Accuracy, Precision, Recall, and F1-score.
RoBERTa gave much better results compared to DistilBERT.
The model predictions also give confidence scores, so users can see how sure the model is.

5. **Deployment**

Built a Gradio interface with a simple text box where users can paste text and get prediction.
Deployed on Hugging Face Spaces so it is publicly available.

Demo link: https://huggingface.co/spaces/akashananddd/capstone_project