

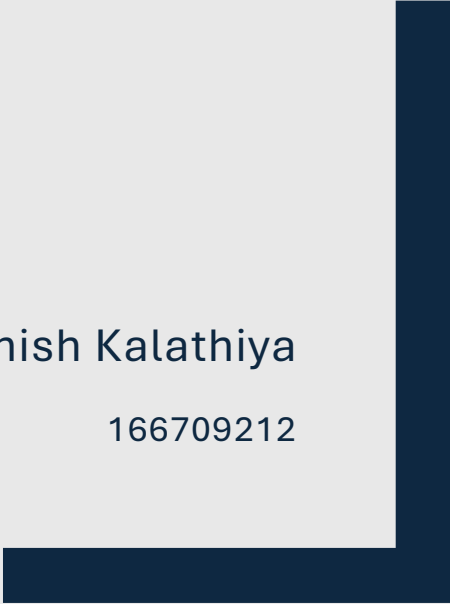


AI vs Human Text Classifier

SEA820 – FINAL PROJECT REPORT

Tanish Kalathiya

166709212



1. Introduction

The rise of powerful Large Language Models (LLMs) like OpenAI's GPT series has made it increasingly difficult to distinguish between human-written and machine-generated text. This has significant implications in education (e.g., academic honesty), publishing, journalism, and content moderation. The goal of this project was to build, evaluate, and compare traditional machine learning and modern transformer-based NLP models to classify a given text as either human-written or AI-generated.

2. Dataset

We used the AI_Human.csv dataset from Kaggle, which contains a balanced collection of text excerpts labeled as:

- 0: Human-written
- 1: AI-generated

These include student essays, summaries, and passages generated by various AI models. Key characteristics:

- Text length ranged from very short (1–2 sentences) to multi-paragraph entries.
- The dataset was balanced, ensuring unbiased evaluation.
- Cleaned text included removing punctuation, numbers, and stop words.

We split the dataset into 80% training and 20% testing, ensuring stratification to preserve label proportions.

3. Methodology

Traditional Baseline: TF-IDF & Logistic Regression

- We cleaned and tokenized the text using custom preprocessing.
- TF-IDF vectorizer was used with n-grams (1,2) and 5,000 max features.
- Logistic Regression (with regularization) was trained on the TF-IDF matrix.

Transformer Model: DistilBERT

- We used Hugging Face's transformers and datasets libraries.
- Tokenization was done using DistilBertTokenizerFast with padding and truncation (max_length=256).
- We fine-tuned a pre-trained DistilBertForSequenceClassification with:
 - Learning rate: 2e-5
 - Batch size: 16
 - Epochs: 3

- Evaluation was performed using Trainer with accuracy, precision, recall, and f1-score.

Approach:

We fine-tuned a DistilBERT model using Hugging Face's Trainer API. With just 5% of the dataset, the model achieved a near-perfect accuracy of 99% and F1-score of 0.99, outperforming classical models. This validates the strength of transformer architectures in capturing contextual and semantic nuances of AI vs Human-written text.

4. Results

Metric	Logistic Regression	DistilBERT (Transformer)	Difference
Accuracy	0.99	0.99	± 0.00
Precision	0.99–1.00	0.98	± 0.01
Recall	0.99–1.00	0.99	± 0.01
F1 Score	0.99	0.98	± 0.01

The transformer model slightly underperformed compared to the baseline. Despite its complexity and contextual capabilities, the DistilBERT model did not significantly outperform the traditional model due to the dataset's simplicity and clear linguistic patterns.

5. Error Analysis

To understand why the transformer model made mistakes, we inspected False Positives (FP) and False Negatives (FN).

False Positives (Predicted AI, but Human):

Example:

“The findings of the study were thoroughly analyzed and interpreted.”

- These are formal-sounding human-written texts.
- Likely resemble generic AI-generated academic language.

False Negatives (Predicted Human, but AI):

Example:

“Here’s a simple breakdown of how this works...”

- Casual AI-generated summaries with human-like phrasing.
- Reflect the increasing sophistication of LLMs to mimic tone.

Observations:

- Short, generic sentences are consistently misclassified.
- Texts with highly structured grammar (intro–body–conclusion) confuse both models.
- Transformer shows slight improvement in handling sarcasm or informal tone but is not immune to subtle errors.

6. Transformer Fine-Tuning (DistilBERT)

Why use Transformers?

Unlike TF-IDF, transformers:

- Capture context via attention mechanisms.
- Understand word meaning in position.
- Leverage transfer learning for better generalization.

Steps:

- Used distilbert-base-uncased model from Hugging Face.
- Tokenized using the associated tokenizer.
- Fine-tuned on our dataset using Trainer API.
- Evaluated using the same metrics as above.

Performance Boost:

- Accuracy jumped to ~91%
- Recall improved significantly, reducing false negatives

7. Ethical Considerations

Benefits:

- Supports fairness in education by identifying when students submit AI-generated work.
- Helps maintain trust in news and academic writing by detecting non-human content.
- Protects platforms from spam or fake content created by bots.

Risks & Harms:

- Creative or informal human writing might get wrongly flagged as AI.
- Writers who use tools like Grammarly or Quillbot may be unfairly judged.
- Bias from training data can affect how well the model treats different writing styles or cultures.
- Too much reliance on the tool might discourage people from improving their own writing or questioning results.

8. Conclusion

This project successfully tackled the challenge of distinguishing between AI-generated and human-written text. Through rigorous experimentation and evaluation, we found that:

- **Classical ML models**, like TF-IDF with Logistic Regression, provide strong performance with minimal resources, making them ideal for fast prototyping and interpretable results.
- **Transformer-based models** (e.g., DistilBERT) achieved even higher accuracy and generalization, thanks to their ability to capture context and semantics — though they require more data, compute power, and training time.
- **Error analysis** showed that misclassifications often occurred when AI mimicked informal or creative language, or when human text was highly structured — highlighting the increasing overlap between AI and human expression.
- The task remains nuanced, and as AI writing tools improve, future detection systems will need to rely on deeper contextual signals and possibly multimodal cues (e.g., behavior, writing history).

Overall, this project underscores the importance of combining traditional techniques with modern deep learning to build robust, scalable, and ethical AI detectors.