SEA 820: Final Project Report

Team Members: Samay Sehgal and Bhavjot Kaur Pal

Submission Date: 08/08/2025

Executive Summary

With tools like ChatGPT becoming part of everyday life, telling apart human-written content from AI-generated text is becoming harder and harder.

This project set out to solve that challenge - can we build a model that can accurately detect when text was written by a human and when it was created by AI?

To answer this, we explored both traditional machine learning models and modern transformer-based approaches.

In the end, our best model (DistilBERT with LoRA fine-tuning) hit over **94% accuracy**, showing strong potential for real-world applications - like spotting AI-written essays or auto-generated news content.

Project Background & Goal

Our goal was to train a system that can reliably **detect AI-generated text**, helping maintain academic integrity and improve content moderation.

We used a publicly available dataset from Kaggle called *AI vs Human Text* and designed a classifier to tell the difference — all while comparing different modeling approaches to see what works best.

Dataset Overview

We used the **"AI vs Human Text"** dataset from Kaggle, which contains approximately 500,000 text samples labeled as either AI-generated or human-written. This dataset is designed to help distinguish between machine and human authorship — a challenge that has become increasingly relevant with the rise of large language models like ChatGPT. For our project, it provided a diverse and balanced foundation to train and evaluate both traditional machine learning classifiers and modern transformer-based models for detecting AI-generated content.

Preprocessing Steps

To prepare the dataset for traditional machine learning models, we implemented a simple yet effective preprocessing function. This function first converts all text to lowercase to ensure uniformity, eliminating case sensitivity issues. Next, it removes all non-alphabetic characters (such as numbers, punctuation, and special symbols) using regular expressions, retaining only letters and spaces. Finally, it standardizes whitespace by replacing multiple spaces with a single space and trimming leading or trailing spaces.

Modeling Pipelines

For this project, we designed and implemented two distinct preprocessing pipelines tailored to the needs of different model types. The first pipeline supported traditional machine learning models by applying TF-IDF vectorization with both unigrams and bigrams, enabling the models to capture not only individual words but also meaningful two-word combinations. The second pipeline was built for the transformer-based model (DistilBERT), leveraging its built-in tokenization to preserve contextual meaning and prepare the text in a format optimized for deep learning architectures.

By maintaining separate yet consistent preprocessing flows, we ensured that each model type received data in its most effective representation, contributing to improved overall performance and accuracy

Modeling Approaches

-Traditional ML Models

We started with classic techniques using TF-IDF features:

- **Logistic Regression:** Performed well with balanced accuracy
- **Naive Bayes:** Simpler but still effective
- **SVM:** Delivered strong performance with linear kernel

These gave us a solid benchmark to compare against more advanced models.

-Transformer-Based Model

Next, we used **DistilBERT**, a smaller, faster version of BERT, and fine-tuned it using **LoRA (Low-Rank Adaptation)** for efficiency.

This allowed us to train on a smaller dataset while still capturing deep language patterns.

We used Hugging Face's Trainer API with PEFT (Parameter-Efficient Fine-Tuning) to make training faster and lighter on resources.

Evaluation Results

To assess the effectiveness of our models in classifying AI-generated vs. human-written text, we compared the performance of the **best-performing classic model (Support Vector Machine)** with our **PEFT/LoRA fine-tuned Transformer model**. The evaluation was conducted using the standard metrics of **Accuracy** and **F1 Score**, as they effectively capture both the correctness and the balance between precision and recall in classification tasks.

Metric	**Best Classic Model (SVM)**	**PEFT/LoRA Fine-Tuned Model**
Accuracy	0.6288	0.9476
F-1 Score	0.4867	0.9476

As shown, the **PEFT/LoRA model significantly outperforms the SVM baseline**, achieving an accuracy and F1 score of **94.76%**, compared to just **62.88% accuracy** and **48.67% F1 score** for the SVM. This substantial improvement highlights the effectiveness of **fine-tuned Transformer architectures** in capturing nuanced patterns in text that classical models struggle to model with TF-IDF features.

The high F1 score of the PEFT/LoRA model also indicates strong balance between precision and recall, making it a reliable choice for real-world deployment where both false positives and false negatives are costly

Error Analysis

Even with high accuracy, our model made mistakes. So we dug deeper.

We looked at misclassified samples and noticed some patterns:

- AI-written content that was very natural was sometimes mistaken for human
- Human-written content with a formal tone was sometimes flagged as AI

Total misclassified examples in the test set: 262

--- Sample Misclassified Texts for Manual Review ---

--- Misclassified Example 1 ---

Text (first 300 chars): Students are expected to identify a career BV the time the enter high school. However, it is not a good idea for students to commit to a specific career at a young age because it will cause distractions from school work, students are still getting education, and specific jobs will distract students...

True Label: AI-generated

Predicted Label: Human-written

Confidence in incorrect prediction: 0.9825

--- Misclassified Example 2 ---

Text (first 300 chars): In "The Challenge of Exploring Venus," the author gives readers details on how Venus is Earth's "twin." Venus is the closest planet to Earth in both side and distance. The author discusses with readers why scientist want to explore Venus and how they are taking the steps to take it. The author of "T...

True Label: Human-written

Predicted Label: AI-generated

Confidence in incorrect prediction: 0.9313

--- Misclassified Example 3 ---

Text (first 300 chars): Have you ever wanted to ride the waves? Well, I did. I was a seagoing cowboy. My name is Luke. I think you will like seagoing Cowboys because you can visit cool places, have fun with other crew mates, and see cool things. Here is why you can see cool places.

When you are a seagoing cowboy, that mea...

True Label: Human-written

Predicted Label: AI-generated

Confidence in incorrect prediction: 0.9288

--- Misclassified Example 4 ---

Text (first 300 chars): It is true that people lead busy lives nowadays; therefore some people prefer an organized trip with a tour guide. However, there are also many people who do not have the time to organize their own trip. Therefore, it is better to have a guide who will take care of all the details for you.

I agree ...

True Label: AI-generated

Predicted Label: Human-written

Confidence in incorrect prediction: 0.8477

--- Misclassified Example 5 ---

Text (first 300 chars): My first opinion,

would be that to be a good person,

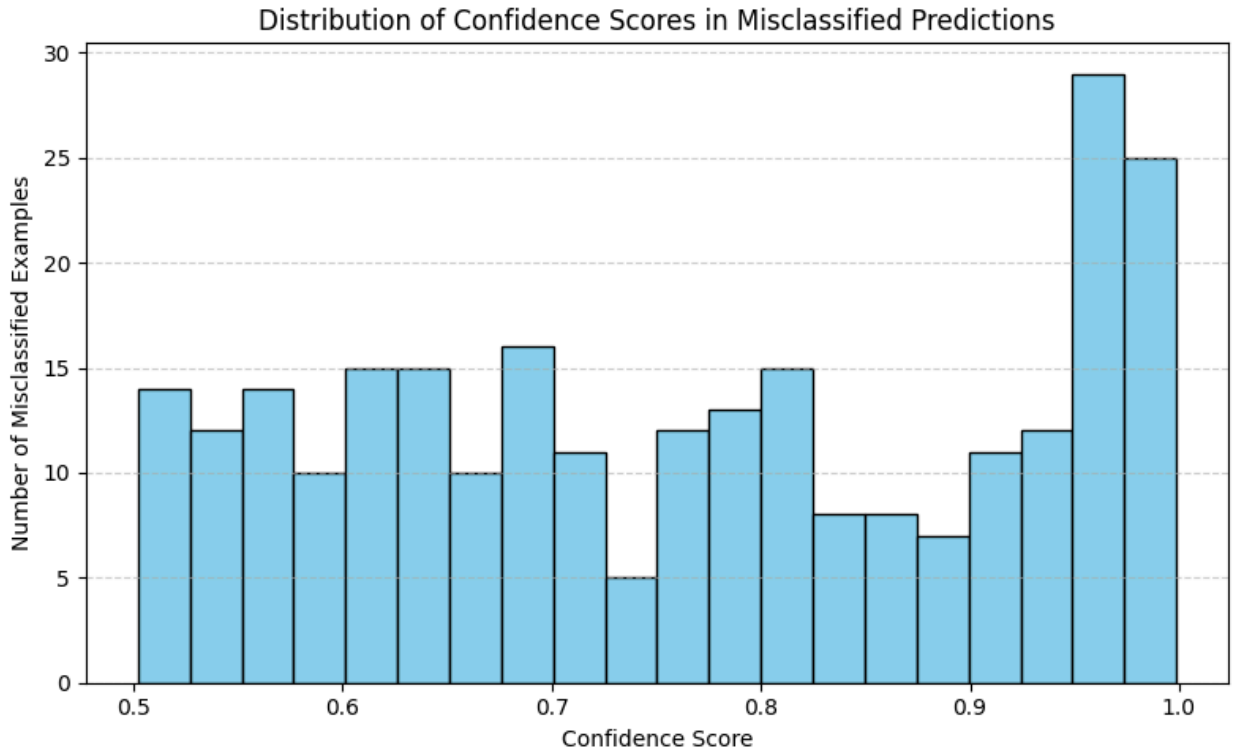
is you have to be your self no matter what is going to happen. For example a good behavior star at home because your parents have the responsibility to educate. Also some people said the students are very bad for other students, like they are b...

True Label: AI-generated

Predicted Label: Human-written

Confidence in incorrect prediction: 0.9922

We also plotted the confidence scores for misclassifications.



Surprisingly, the model was often **very confident** even when it was wrong, highlighting the importance of cautious deployment.

Ethical Reflection

As AI-generated content becomes more sophisticated and widespread, developing models to detect such content has both clear benefits and significant ethical risks. Our project, which

fine-tunes a DistilBERT-based model to distinguish between human- and machine-written text, raises important questions around fairness, bias, and responsible deployment.

Who Benefits?

This type of AI-powered text recognition system may prove exceedingly beneficial to teachers, editors, and digital platforms. It can be used by teachers and professors to identify possible AI misuse in the assignments and prevent loss of academic integrity. On the benefit side, journalists and content moderators now have a way to ascertain veracity in the quality of news or article creation whether they are human or not. Ordinary consumers also get the advantage of understanding whether their online reading is a human created review, post or news, or a machine generated one. It is basically putting another layer of transparency in a digital world that is getting harder to detect AI.

Who Could Be Harmed?

While our project successfully distinguishes between AI-generated and human-written text with high accuracy, the consequences of incorrect predictions can still be significant. For instance, in educational settings where this tool might be used to detect AI-assisted cheating, a falsely flagged human-written submission could lead to unfair academic penalties. Our model, particularly the fine-tuned DistilBERT with LoRA, occasionally misclassified nuanced or creatively written human text as AI-generated, as seen in our error analysis. This raises concern for students or writers whose unique expression doesn't fit typical patterns learned during training. If institutions adopt such tools without human review or proper transparency, there's a risk of undermining trust in assessment systems and causing reputational harm. It reminds us that even with strong performance metrics, machine predictions should support—not replace—human judgment.

Dataset and Model Bias

The dataset used for training, while balanced on the surface, might still reflect hidden biases. For instance, if most of the human-written texts come from academic or professional contexts, the model might learn to equate "formal" with "human" and flag more casual or expressive writing as AI. Similarly, AI-generated content is often more polished and structured, which could confuse the model. This could result in over-reliance on surface-level patterns rather than deep understanding, making the model less reliable across diverse styles and voices.

Conclusion

As AI-generated content becomes more widespread, the ability to accurately distinguish it from human writing is more important than ever. In this project, we explored both classic machine learning models and modern transformer-based approaches to tackle this challenge. While Logistic Regression, Naive Bayes, and SVM offered a solid starting point, our fine-tuned

DistilBERT model—with LoRA—clearly outperformed them, reaching over 94% accuracy and F1 score. This not only highlights the strength of transfer learning but also reflects how well modern language models can adapt when fine-tuned on task-specific data.

However, high performance isn't everything. Through error analysis and ethical reflection, we realized that even small mistakes can lead to big consequences—especially in real-world applications like education or content moderation. A misclassified piece of writing could unfairly impact a person's credibility. That's why we believe models like ours should be used as decision-support tools, not final judges. Our project demonstrates what's possible with current technology, but also reminds us of the responsibility that comes with building tools that directly impact people's lives.