MODEL FORGE

DOMAIN : NATURAL LANGUAGE PROCESSING

Natural Language Processing

We chose NLP because language is the richest signal for detecting writing origin. Human writing carries emotional nuance, stylistic inconsistency, and personal voice — patterns that both classical ML and transformer-based models can learn to distinguish from the uniform, structured output of AI language models.

Text Classification

Transformers

Binary Labels

105K Corpus

// NLP · BINARY CLASSIFICATION

AI vs Human
Text Classifier

Binary Classification · 105K Dataset · Human=1 / AI=0

TEAM MEMBERS

V. Sree Kirthana
ML Engineer

Sohail Azain
Data Scientist

Sakshi Yadav
NLP Researcher

JBIET · B.Tech 2nd Year · 2024–25

Project Objective

Build a binary text classification model that automatically detects whether a given piece of text was written by a Human (label: 1) or generated by an AI system (label: 0) — trained on a large-scale dataset of 105,000 text samples.

Dataset

105K labeled text samples
Human=1 / AI=0 labels

Models

Compare ML & DL approaches
Logistic Reg → SVM → DistilBERT

Target

Maximize classification accuracy on unseen data

Result

Achieved 93% final accuracy with DistilBERT transformer

Logistic Regression

How it Works

Feature Extraction:
Text vectorized using TF-IDF (Term Frequency–Inverse Document Frequency)

Algorithm:
Applies sigmoid function to predict probability of class (Human or AI)

Decision Boundary:
If P(Human) > 0.5 → Label = 1 (Human), else Label = 0 (AI)

Advantages:
Fast to train, interpretable, strong baseline for text classification

ACCURACY

89%

on test set

Performance Metrics

Metric	Human (1)	AI (0)
Precision	88%	90%
Recall	91%	87%
F1-Score	89%	88%

04 / MODEL 2

Linear SVM

How it Works

Feature Extraction:
TF-IDF vectorization converts raw text into numerical feature vectors

Core Idea:
Finds the optimal hyperplane that maximizes margin between Human and AI classes

Why Linear Kernel:
Text data in high-dim TF-IDF space is linearly separable — linear kernel is efficient and effective

Advantage over LR:
Better margin maximization → improved generalization on unseen text data

ACCURACY

91%

+2% vs LR

Performance Metrics

Metric	Human (1)	AI (0)
Precision	90%	92%
Recall	93%	89%
F1-Score	91%	90%

DistilBERT

Transformer Model — Best Result

What is DistilBERT?

Architecture:
Distilled (compressed) version of BERT — 66M parameters, 40% smaller, 60% faster

Tokenization:
WordPiece tokenizer captures subword context; handles rare words and typos

Attention Mechanism:
Self-attention layers understand full context of each word in the sentence — not just adjacent words

Fine-tuning:
Pre-trained on massive corpus, then fine-tuned on our 105K dataset with a classification head

FINAL ACCURACY

93%

+4% vs SVM · Best Model

Performance Metrics

Metric	Human (1)	AI (0)
Precision	93%	94%
Recall	94%	92%
F1-Score	93%	93%

Why Classification Works

01 **Raw Text Input —** 105K text samples fed into the pipeline. Each labeled Human (1) or AI (0)

02 **Preprocessing —** Lowercasing, stopword removal, punctuation cleaning, tokenization

03 **Feature Repr. —** ML: TF-IDF sparse vectors
DL: DistilBERT contextual token embeddings

04 **Model Training —** Model learns patterns that differentiate human writing style vs AI generation

05 **Prediction —** Output probability → threshold at 0.5 → final label (Human or AI)

Tools & Libraries Used

Platform

Google Colab
Free GPU/TPU access in the cloud, Jupyter notebook interface, no local setup needed

Dataset Source

Kaggle — AI vs Human Text Dataset
105,000 rows · 2 columns: text, label
Split: 80% Train / 10% Val / 10% Test

Python 3.x Language	**NumPy** Numerical	**Pandas** Data	**Scikit-learn** ML Models
NLTK NLP Preprocess	**re / string** Text Cleaning	**PyTorch** Deep Learning	**Transformers** Hugging Face
DistilBERT Pre-trained	**Matplotlib** Visualization	**Seaborn** Plots	**TensorFlow** Alt DL

Conclusion

This project successfully demonstrated that AI-generated text can be distinguished from human-written text using a progressive ML pipeline — culminating in a 93% accuracy with DistilBERT on a 105,000-sample dataset.

01

Classical ML is a strong start — Logistic Regression (89%) and Linear SVM (91%) proved that TF-IDF features carry significant discriminatory power for this task.

02

Transformers outperform ML — DistilBERT's contextual embeddings captured deep semantic patterns, pushing accuracy to 93% — a 4% gain over SVM.

03

Scale matters — Training on 105K samples provided enough diversity for the models to generalize well across different text styles and AI generators.

89%	→	91%	→	93%	✓
Logistic Reg		Linear SVM		DistilBERT	Best Model