SEA 820 NLP Final Project: Detecting AI-Generated Text

Course: SEA820
Abhi Patel
Arnav Nigam

Motivation & Background

- Rise of LLMs (e.g., ChatGPT, Gemini) harder to distinguish between human vs AI text
- Real-world impact: academic integrity, misinformation, moderation

Project Goal and Overview

The goal of this project is to build, evaluate, and compare different NLP models for the task of classifying a given text as either "human-written" or "AI-generated."

Dataset Overview

- Source: Kaggle – *AI vs Human Text*
- ~480k samples
- Each sample is labeled as either 0 for human or 1 for AI.
- Used entirely all of the dataset for classical model in part 1.
- Downsampled to **100,000 in part 2** for resource efficiency & fast iteration.

Data Preprocessing

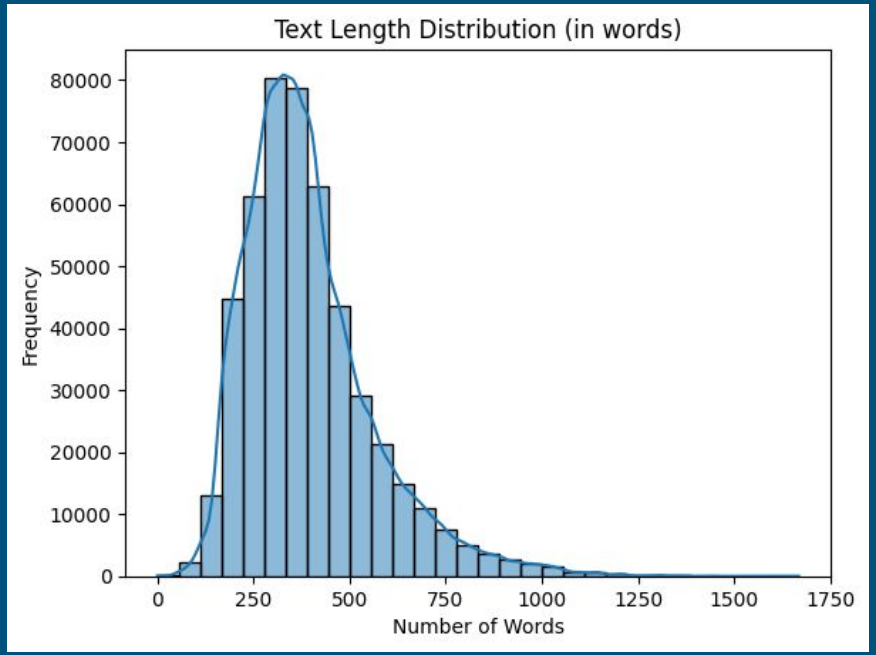
Our pipeline involved the following steps:

- Lowercasing,
- Punctuation/digit removal,
- Stopword removal,
- Lemmatization (WordNet).
- Tokenization(NLTK)

Quality Control: Removed empty texts, applied consistent preprocessing, preserved original text length stats.

Text size and occurence

mean	393.096214
std	168.593328
min	0.000000
25%	278.000000
50%	363.000000
75%	471.000000
max	1668.000000



Logistic Regression

Accuracy: 99.53% - Only ~0.47% of predictions are wrong

Error Rate: ~0.47% on the validation set

Precision: 99.52% - Very few false positives

Recall: 99.48% - Captures nearly all true positives

Observation:

- Performs exceptionally well on internal test data
- Slightly complex compared to Naive Bayes due to parameter optimization and iterative convergence
- More computationally intensive for large datasets

Naive Bayes

Accuracy: 97.07% ~2.93% error rate

Precision: 97.07%

Recall: 96.65%

Observation:

- Lower scores than Logistic Regression but still high accuracy
- Extremely fast to train (milliseconds) — important for large datasets
- Performs consistently even with sparse TF-IDF features
- Simpler to implement and interpret — ideal as a **baseline** before moving to complex transformers

Final Baseline Model

Baseline Justification:

- In Phase 2, the baseline model is not meant to be the absolute best — it's a reference point to prove transformers can outperform classical ML
- Naive Bayes' speed and simplicity make it the perfect benchmark for relative improvement measurement
- Produced stable results across multiple random samples
- Demonstrated robustness to text distribution shifts in smaller test subsets

Example from notebook:

- On a quick 20k random subset:
- Naive Bayes: ~97% accuracy in under 2 seconds training time
- Logistic Regression: ~99% accuracy but took 30–60 seconds training

Phase 2 - Transformer Models

Goal: Beat Phase 1 baseline model.

Chosen baseline: Naive Bayes, accuracy 97.07%

Models Used:

- DistilBERT (smaller, faster)
- RoBERTa (larger, more powerful)

Tools: HuggingFace Transformers + Datasets

DistilBERT Fine-Tuning

- **Hyperparameters**:
  - LR: 2e-5, Epochs: 3, Batch size: 16
  - Optimized for small model, avoid overfitting

- **Results** (on 10K validation set):
  - Accuracy: **99.67%**
  - Precision: 99.71%, Recall: 99.42%, F1: **99.57%**

- Error analysis: 33 misclassified samples, often borderline cases

RoBERTa Fine-Tuning

- Hyperparameters:
  - LR: 3e-5, Epochs: 2, Batch size: 32

- Larger model, fewer epochs to prevent overfitting
- Results (on 10K validation set):
  - Accuracy: 99.71%
  - F1: 99.71%

- Slight improvement over DistilBERT ~0.04% accuracy change

Model Comparisons

Model	Accuracy	F1 Score
Naive Bayes	97.07%	99.57%
DistillBERT	99.67%	99.57%
RoBERTa	99.71%	99.71%

Evaluation on External Samples

Human-written samples:

- All 5 misclassified as AI
- Confidence > 99.8%
  Suggests model is very sensitive to formal tone or structured grammar

AI-generated samples:

- All 5 correctly classified with **>99.9%** confidence
- Highlights strength in AI detection

Error Analysis

False Positives (Human texts being tagged as AI):

- Formal, academic-style human writing

False Negatives (AI text being tagged as Human): Very rare

Possible causes:

- Overfitting to stylistic patterns
- Lack of diversity in training samples

Ethical Considerations

Bias: Model may flag formal non-native English as AI

Fairness: Penalizing students or writers with a structured style

Use Cases:

- Good: moderation, verification tools
- Bad: automated punishment, surveillance

Lessons & Takeaways

- Transfer learning significantly outperforms classical methods
- Even small models (DistilBERT) can deliver >99% accuracy
- Careful hyperparameter tuning + error analysis is key
- Human-like text can still fool the model → needs further refinement
- Try Prompt Tuning or LoRA (PEFT methods)