AI-Generated Text Detection

**Team Members:** Abhi Patel & Arnav Nigam

**Course:** SEA 820 - Natural Language Processing

**Date:** 8th August 2024

Table of Contents

[Abstract 1](#_Toc205539654)

[Introduction 1](#_Toc205539655)

[Dataset 1](#_Toc205539656)

[Methodology 3](#_Toc205539657)

[Results and Analysis 4](#_Toc205539658)

[Ethical Considerations and Societal Impact 8](#_Toc205539659)

[Challenges faced 9](#_Toc205539660)

[Conclusion 9](#_Toc205539661)

## Abstract

The rapid evolution of Large Language Models (LLMs) such as GPT and Gemini has significantly blurred the line between human-written and machine-generated text. This poses challenges for academic integrity, online content moderation, and combating misinformation. This project investigates the use of both classical and Transformer-based models for detecting AI-generated text, using the publicly available AI vs Human Text dataset. Phase 1 established a baseline using a Naive Bayes classifier with TF-IDF features, achieving an accuracy of 97.07%. In Phase 2, we fine-tuned two pre-trained Transformer architectures namely: DistilBERT and RoBERTa using the Hugging Face Trainer API. DistilBERT achieved 99.67% accuracy, while RoBERTa achieved 99.71%, both outperforming the baseline by a significant margin. Comprehensive evaluation, including precision, recall, and F1-score, revealed exceptional classification performance with minimal misclassification. However, external testing highlighted high sensitivity to formal human writing, raising potential fairness and bias concerns. This work demonstrates that fine-tuned Transformer models can reliably detect AI-generated text, though ethical considerations remain critical before deployment.

## Introduction

The emergence of advanced LLMs and their derivatives have transformed content generation across industries. These systems can produce coherent, contextually relevant, and stylistically diverse text that often rivals human writing. While this capability presents opportunities for automation, creativity, and efficiency, it also introduces risk such as academic dishonesty, misinformation dissemination, and challenges to authorship verification.

This project aims to build, evaluate, and compare models for classifying text as human-written or AI-generated. We employ a two-phase approach: first, establishing a classical baseline with a Naive Bayes classifier and TF-IDF features; second, leveraging transfer learning by fine-tuning Transformer models, DistilBERT and RoBERTa, on the AI vs Human Text dataset. Beyond performance evaluation, we conduct a detailed error analysis, assess model robustness on external samples, and reflect on the ethical implications of deploying such technology.

## Dataset

The "AI vs Human Text Language" dataset from Kaggle was used for the text classification task, containing 487,235 text samples with binary labels (0: human, 1: AI-generated). There is a significant imbalance in the dataset in that 65% of the texts were human produced while the remaining 35% were AI productions. The category mainly consists of student essays and some educational materials.

A graph of a bar chart

AI-generated content may be incorrect.

The dataset had a mean of approximately 390 words per samples and a max of 1668 words. However most of it lied between 250 – 500 words. See figure below:

A graph of a number of words

AI-generated content may be incorrect.

## Methodology

### Data Preprocessing

For our data preprocessing pipeline, we started off with spacy. However, realised that the google colab memory is being overused. This made us switch to NLTK which is lighter and faster compared to Spacy and is highly compatible for this task.

1. **Text cleaning involved 5 steps:**
   * Lowercasing,
   * punctuation/digit removal,
   * stopword removal,
   * lemmatization (WordNet).
   * Tokenization(NLTK)
2. **Quality Control:** Removed empty texts, applied consistent preprocessing, preserved original text length stats.
3. **Down sampling**: Stratified sampling to 100,000 for resource efficiency and faster iteration in phase 2 of our project.

### Baseline Models (Classical ML)

1. **Feature Engineering:**

TF-IDF with max\_features=15000, min\_df=5, max\_df=0.85, ngram\_range=(1,2), sublinear\_tf=True.

1. **Models:**

* **Logistic Regression:** class\_weight='balanced', max\_iter=1000, L2 regularization.
* **Multinomial Naive Bayes:** Optimized for TF-IDF, no tuning.

### Transformer Models (Fine-Tuning)

1. **Architectures:**

* **DistilBERT-base-uncased:** 6 layers, 768 hidden units, 12 heads, 66M params.
* **RoBERTa-base:** 12 layers, 768 hidden units, 12 heads, 125M params.

1. **Training Setup:**

Balanced 100k dataset, 80/20 split, max length 256, dynamic padding/truncation.

* **DistilBERT:** LR=2e-5, batch=16, epochs=3, weight decay=0.01.
* **RoBERTa:** LR=3e-5, batch=32, epochs=2, weight decay=0.05.

1. **Evaluation**

* **Metrics:** Accuracy, Precision, Recall, F1, ROC-AUC.
* **Validation:**
  + Stratified upto 20% test set.
  + External test on real human-written samples.
  + Side-by-side model comparison.

## Results and Analysis

Comparison of all models

| Model | Accuracy | Precision | Recall | F1-Score | ROC-AUC |
| --- | --- | --- | --- | --- | --- |
| Logistic Regression | 99.53% | 99.52% | 99.48% | 99.50% | 0.999 |
| Naive Bayes | 97.07% | 97.07% | 96.65% | 96.85% | 0.985 |
| DistilBERT | 99.72% | 99.71% | 99.61% | 99.72% | 0.999 |
| RoBERTa | 99.66% | 99.71% | 99.42% | 99.66% | 0.999 |

**Key Observations**

* **Transformers have performed the best with** DistilBERT and RoBERTa clearly outperforming the classical models, showing the strength of modern NLP architectures for this task.
* DistilBERT slightly surpasses RoBERTa’s performance, which shows that a lighter model can still deliver better results.
* Logistic Regression comes surprisingly close to the transformers, suggesting that the dataset suits traditional text classification methods well.
* All models scored above 97%, indicating the dataset is relatively clean and the classification task is straightforward.

## Error Analysis

**Internal Test Set Errors**

* Logistic Regression: 204 human texts misclassified as AI; 258 AI texts misclassified as human (0.47% error rate).
* Naive Bayes: 1,037 humans → AI, 1,814 AI → human (2.93% error rate).
* Transformers: Near-perfect performance. DistilBERT (26 errors/10k), RoBERTa (33 errors/10k).

Have a look at the confusion matrix below for each model after undergoing an evaluation test for easier navigation:

**Logistic Regression:**

**A blue squares with white text

AI-generated content may be incorrect.**

**Naïve Bayes:**

**A green and white chart

AI-generated content may be incorrect.**

**RoBERTa:**  
A blue squares with white text

AI-generated content may be incorrect.

A blue squares with white text

AI-generated content may be incorrect.**DistilBERT-uncased:**

**When tested on unseen human-written samples, overfitting became clear:**

* **Human Texts:** Logistic Regression (40% accuracy), Naive Bayes (100%), DistilBERT (0%), RoBERTa (0%).
* **AI Texts:** All models classified correctly with 99%+ confidence.

**Error Patterns**

* Shows a heavy burden of writing styles found in a dataset.
* Overfitting mostly to educational themes.
* Preference for a particular length of text.
* Decision-making based on words belonging to a specific domain.

**Specific Examples of Misclassification:**

**Human Text (Misclassified as AI)***:*"The school times were the best, we had the opportunity to hang around, be lazy and don't do our assignments, I wish I was still in school."

**Analysis:** Contains informal language and personal emotions which should be classified as human

**AI Text (Correctly Classified)***:*"The implications of artificial intelligence in healthcare are far-reaching, promising better diagnosis and treatment outcomes through machine learning."

**Analysis:** Formal, structured language with technical terminology - clear AI-generated characteristic

### Model Behavior Insights

**Feature Importance**

* **TF-IDF:** Educational terms and bigrams were strong indicators.
* **Transformers:** Focused on sentence structure, punctuation, and topic-specific words, but often ignored human-like stylistic cues.

**Confidence Calibration**

* Transformers were **overconfident** (99%+ on most predictions, even ambiguous ones).
* Naive Bayes produced more estimations which were balanced on probability.

## Ethical Considerations and Societal Impact

### Potential Benefits

* **Academic Integrity:** To help detecting AI-generated essays, protect assessment validity, and observe class honesty policies.
* **Content Moderation:** To track misinformation spread using AIs, block automatic spams, and stop disinformation campaigns.
* **Research Value:** To study the generation of AI texts, make detection methods better, and learn more about the distinction between human and AI writing.

### Potential Harms and Risks

Bias and Discrimination:

* It penalizes the non-native English speakers who may write things considered formal or "perfect," with unusual grammar patterns, or the absolute minimum use of idioms.
* It may classify legitimate genres technically, academically, and creatively as AI compositions.

Misuse and Abuse:

* **Academic Surveillance:** Too much trust is put in AI to replace human judgment, thus accusing some students and replacing human judgement.
* **Content Censorship:** Inappropriate removal of legitimate works.

Ethical Questions Arising from Our Study

* Overfitting: While the models performed well in test data, poor results in situations concerning real-world data gave a false impression of reliability during high-stake circumstances.
* Dataset Bias: Because of its heavy concentration on educational texts, it causes domain-specific overfitting and weakens performance in terms of other content.
* Overconfidence: Transformer models often gave 99%+ confidence even when wrong, which can be misleading for decision-makers.

## Challenges faced

* Dataset Size and Computational Constraints

The original AI vs Human Text dataset contained over 480,000 samples. While larger datasets can improve generalization, fine-tuning Transformer models on the entire dataset exceeded our available computational resources and Colab runtime limits. To address this, we down sampled to 100,000 stratified samples, ensuring representative coverage while enabling timely training without frequent session timeouts.

* Tokenization and Label Formatting Issues

Early in the fine-tuning process, we encountered multiple errors related to label type mismatches. Hugging Face’s Trainer API requires integer-encoded labels, but our preprocessed dataset initially retained float or string label formats. This caused runtime failures, particularly during GPU execution. We resolved this by explicitly casting labels to ClassLabel objects in the Hugging Face Datasets library, ensuring proper compatibility.

* Overfitting Risks

Initial experiments with DistilBERT revealed signs of overfitting after only a few epochs, given the high baseline accuracy of classical models in Phase 1. This required careful adjustment of hyperparameters, including reducing the number of epochs, modifying batch sizes, and adding weight decay for RoBERTa to promote generalization.

## Conclusion

This study shows a clear gap between how well models perform in the lab and how they behave in the real world. On our internal test set, transformer models like DistilBERT and RoBERTa delivered almost perfect results. But when faced with new, real-world human-written text, their accuracy dropped dramatically. This reminds us that strong test scores don’t always mean a model will work well outside its training environment.

Surprisingly, Naive Bayes, a much simpler model, handled external samples better, even though its test accuracy was lower. This suggests that in some situations, less complex methods can be more reliable than state-of-the-art approaches.

The overfitting we saw in transformer models isn’t just a technical issue rather an ethical one. If such systems were deployed as-is, they could wrongly accuse people, reinforce biases, and create a false sense of trust in their accuracy. Part of the problem lies in the dataset itself, which was heavily focused on educational writing and lacked diversity in topics and styles.

In the end, detecting AI-generated text is not only a technical challenge but also a question of fairness, transparency, and responsibility. While today’s models are impressive, they are not ready for high stakes use without better training data, stronger evaluation methods, and meaningful human oversight.

### References:

All used code and explanations for this project were used from our weekly lectures and weekly labs, mainly Week 1-3 and Week 7-11