SEA 820 NLP Final Project Report: Detecting AI-Generated Text

Author: Arhaam Khan

ID: 162087217

Program: BSA

Contents

Overview

The widespread adoption of Large Language Models (LLMs) like GPT has created a new challenge: distinguishing between human-written and machine-generated text. This has significant implications for academic integrity, online content moderation, and misinformation detection.

Goal

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

The Dataset

1. **Source**: AI vs Human Text dataset from Kaggle.
2. This dataset is specifically curated for this classification challenge.
3. The dataset includes student essays and other text forms, making it relevant and challenging.
4. Each text sample is labeled as either 0 (human) or 1 (AI-generated).

Methodology

1. **Data Exploration & Preprocessing:**
   - Downloaded and loaded the dataset.
   - Performed a thorough exploratory data analysis (EDA). Analyze text length, vocabulary, and class distribution.
   - Created a robust data preprocessing pipeline.
   - Decided on tokenization, cleaning, and how to handle text lengths.

2. **Classic Model (TF-IDF + Logistic Regression)**
   - Preprocessing: Tokenization, stopword removal, punctuation removal, lemmatization.
   - Feature extraction: TF-IDF vectorization.
   - Classifier: Logistic Regression (*scikit-learn*).
   - Experiments with **full dataset** and **5K stratified subset.**

3. **Transformer Model (Hugging Face, DistilBERT)**
   - Tokenization using *AutoTokenizer* from Hugging Face.
   - Fine-tuned for binary classification using *Trainer* API.

- Used stratified **5K sample** for faster training.

Results

Model	Accuracy	Precision	Recall	F1-Score
TF-IDF + Logistic Regression (Full Dataset)	0.95	0.97	0.89	0.93
TF-IDF + Logistic Regression (5K Subset)	0.87	1.00	0.65	0.79
Hugging Face, DistilBERT(5K Subset)	0.97	0.96	0.96	0.96

Analysis

(source: https://developers.google.com/machine-learning/crash-course/classification/accuracy-precision-recall)

1. **TF-IDF + Logistic Regression (Full Dataset)**

- Achieved an accuracy of 0.95, with high precision (0.97) and good recall (0.89), resulting in an F1-score of 0.93.
- The strong recall indicates that the model was able to correctly identify most AI-generated samples while maintaining a **low false positive rate**.

2. **TF-IDF + Logistic Regression (5K Stratified Subset)**

- Achieved an accuracy of 0.87, with perfect precision (1.00) but lower recall (0.65), resulting in an F1-score of 0.79.
- This suggests that while the model was highly confident when predicting AI-generated text **(no false positives)**, it missed a significant portion of AI-generated samples **(false negatives)**, likely due to the reduced size of training dataset.

3. **DistilBERT (Hugging Face, 5K Stratified Subset)**

- Achieved the highest accuracy at 0.97, with balanced precision (0.96) and recall (0.96), resulting in an F1-score of 0.96.

- This balance indicates that the model was equally effective in minimizing both **false positives and false negatives**, despite being trained on the same reduced dataset size as the smaller TF-IDF subset model.

Ethical Considerations

AI detection tools are designed to identify AI-generated content, particularly in educational settings, with the goal of upholding academic integrity. However, any false positives associated with these models come with potentially serious consequences for accused students.

From research, in a Bloomberg test of two AI detectors (GPTZero and CopyLeaks), false positive rates were found to be 1–2% when a sample of 500 essays was run through the checkers. Detectors can also miss AI-generated writing, marking it as human-generated. With respect to false positives, even a small error rate can add up. “If a typical first-year student writes 10 essays, and there are 2.235 million first-time degree-seeking college students in the U.S., that would add up to 22.35 million essays. If the false positive rate were 1%, then 223,500 essays could be falsely flagged as AI-generated (assuming all were written by humans).” This is a substantial number.

In addition to the potential psychological impacts on students, there are also material consequences, including academic penalties, loss of scholarships, and damage to future opportunities. Likewise, with respect to non-native English speakers, use of such models for detection might have them falsely flagged as producing AI-generated text. This is because their writing style may appear extremely formal, simplistic or syntactically different when compared to native English writing. Thus, unintentionally matching patterns that the model associates with AI-generated content. Such bias could further disadvantage students already navigating language barriers.

In the end, AI detectors and their apparent advantages can be outweighed by serious disadvantages. Students may face long-term repercussions from false positives, unjustified academic sanctions, and emotional distress. To overcome these obstacles, educators should prioritize equity and fairness by becoming more knowledgeable about AI and encouraging students to use AI tools responsibly, critically, and thoughtfully. By implementing these measures, we can address the challenges posed by generative AI while creating a fair and inclusive learning environment for all students.

(**Source** : https://citl.news.niu.edu/2024/12/12/ai-detectors-an-ethical-minefield/#:~:text=AI%20detectors%20are%20often%20marketed,long%2Dterm%20consequences%20for%20students.)

Conclusion

- **Speed and Resource Comparison:** Logistic Regression is extremely fast to train (within seconds) and can process thousands of documents per second on a CPU. DistilBERT, while more accurate, is slower to train (tens of minutes even with a GPU for a few epochs) and slower during inference due to its computational complexity. For large-scale deployments (e.g., millions of requests), a smaller transformer model or a traditional approach like Logistic Regression might be more practical unless GPU servers are available.

- **When to Use Which:** For limited data or when a quick baseline is needed, TF-IDF + Logistic Regression or Naive Bayes are efficient starting points. When the highest possible accuracy is required and resources allow, fine-tuning a transformer model is the state-of-the-art choice. A middle ground is using pre-trained transformers to generate embeddings, then training a simpler model on those embeddings, which can yield good accuracy without the heavy cost of full fine-tuning.

- **Overall:** This project demonstrated that both approaches can achieve strong performance, but the transformer-based model generally leads to better accuracy. Traditional models emphasize feature engineering, while BERT-based models rely on leveraging pre-trained architectures and fine-tuning. Each method has its strengths, and the choice depends on the specific use case, available resources, and performance requirements.

-x-