ernest.oseiamoafo@stud.uni-goettingen.de

Ernest Osei Amoafo

AI Content Detection

Comparing Watermarking and Transformer Methods in Terms of Accuracy and Computational Efficiency

Table of Contents

Motivation

- AI Detectors are "Pseudoscience" and incompetent!

Ref: Janelle Shane: "Don't use AI detectors for anything important."

Problem Outline

- AI tools like GPT generate human-like text, complicating the distinction.
- **"Perfect" writing is penalized.**
- The arms race between generation and detection is unwinnable.
- **Challenges:**
  - Plagiarism detection and authenticity of content.
  - **High-Performance Data Analytics (HPDA):** Processing vast datasets requires scalable solutions.

Goal

Objective

Compare detection methods for **Accuracy** and **Efficiency**.

- **In HPDA:** Need scalable methods to handle big data without high computational cost.
- **Focus:** Text-based detection, as datasets are abundant.

State-of-the-Art Methods

Main Categories:

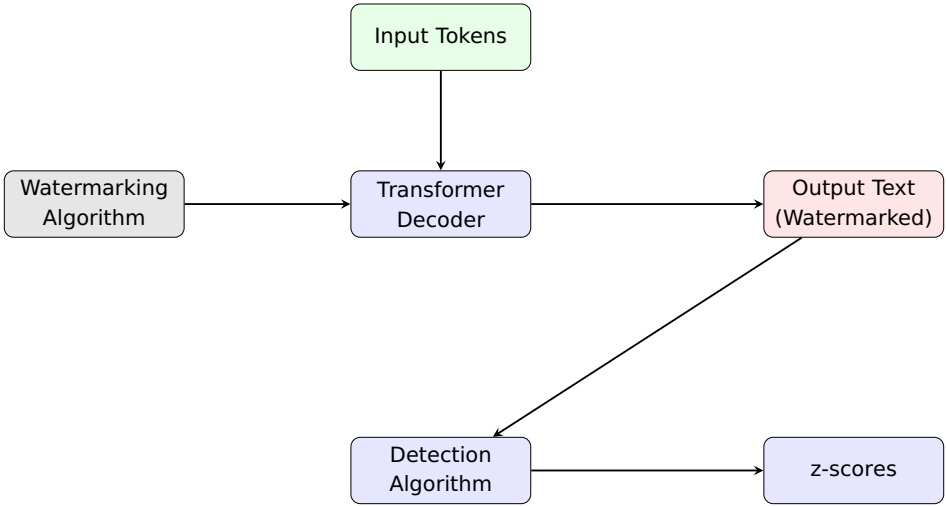
1. **Statistical/Stylistic Analysis:** Detects patterns like repetition or burstiness.
2. **Deep Learning:** Transformers (e.g., BERT) identify inconsistencies.
3. **Watermarking:** Embeds subtle signals in AI outputs for detection.
4. **Hybrids:** Combine methods for better robustness.

Ref: Cao (2025): Watermarking for AI Content Detection; Khan et al. (2025): Identifying AI Content using DistilBERT

Method 1: Watermarking

- **Approach:** Embed probabilistic "watermarks" (e.g., token biases) during generation; detect via statistical tests.
- **Implementation:** Python-based statistical analysis on token distributions (Green-Red List).

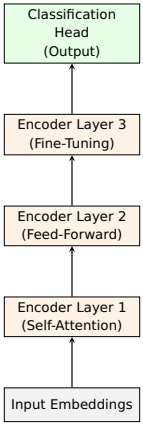
Watermarking Workflow



Method 2: Transformer Model

- **Approach:** Fine-tune a BERT variant (RoBERTa) on labeled texts to classify AI vs. Human.
- **Focus:** Feature extraction for generation patterns (e.g., inconsistencies).

RoBERTa Transformer Architecture



Selected Dataset

- **Source:** Kaggle - AI vs Human Text
- **Size:** 2,000 samples (Stratified subset)
- **Split:** 70% train, 15% validation, 15% test

Ref: https://www.kaggle.com/datasets/shanegerami/ai-vs-human-text

Implementation Overview

- **Repository:** gitlab.gwdg.de/ernest.oseiamoafo/ai-text-detection
- **Tools:** Python (scikit-learn, Hugging Face Transformers).
- **Setup:** Preprocess texts, train/test split (70/15/15).
- **HPDA:** Parallel processing with PyTorch for large batches.
- **Dependencies:** `numpy`, `pandas`, `torch`, `transformers`.

Implementation: Watermarking

- **Algorithm:** Green-Red list method.
- **Parameters:** $\delta = 2.0$, $\gamma = 0.5$
- **Detection:** Z-test with threshold 4.0.

Ref: Kirchenbauer et al., 2023.

Implementation: Transformer Model

- **Model:** RoBERTa-base
- **Training:** 3 epochs, batch size 8
- **Fine-tuning:** Binary classification on Human vs AI text.

Experimental Setup

- **Hardware:** Standard CPU.
- **Metrics:** Accuracy, Throughput, F1-score, Training time.
- **HPDA Metrics:** Inference Time, Throughput (Samples/Sec).
- **Variations:** Scaled dataset sizes to test HPDA suitability.

Implementation Pipeline

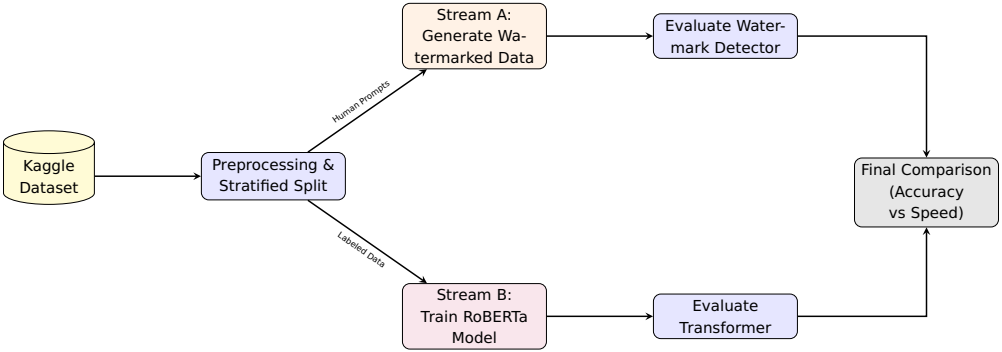
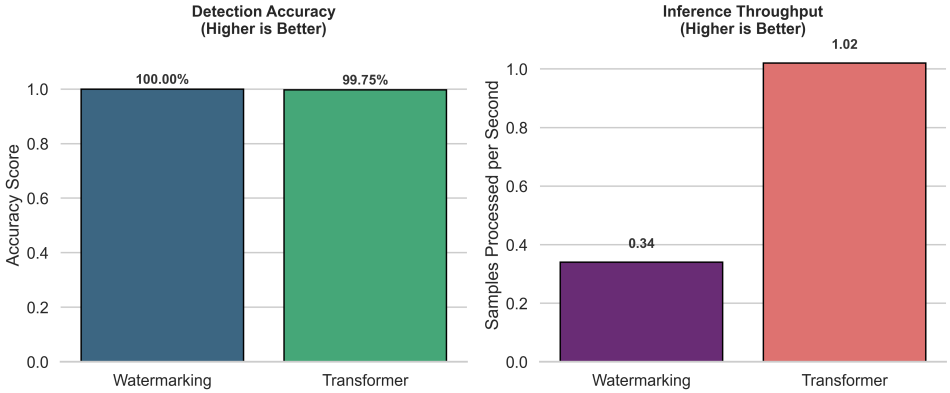


Figure: Dual-stream pipeline ensuring valid testing for both active (Watermark) and passive (Transformer) methods.

Results: Accuracy & Efficiency

Metric	**Watermarking**	**Transformer**
Accuracy	**100%**	99.75%
F1-Score	**1.000**	0.997
Inference Speed	0.34 samples/sec	**1.02 samples/sec**
Training Time	**0 hours** (Zero-shot)	10.5 hours (3 epochs)

Visual Comparison



Left: Accuracy (Both High) | Right: Throughput (Transformer is 3x Faster)

Key Findings

1. **Accuracy:** Both methods are excellent (100% vs 99.75%).
2. **Efficiency Trade-off:**
   - **Watermarking:** Zero training cost (Instant deployment), but slower per-sample inference.
   - **Transformer:** Massive upfront training cost (10+ hours), but **3x faster** inference throughput.
3. **HPDA Implication:** Use Watermarking for agility; Use Transformers for high-volume batch processing.

Conclusion

- Transformers are highly effective (99.75%), but they incur a huge **initial training time**.
- Watermarking offers an immediate, **highly accurate** solution, but lacks the high inference throughput of optimized neural networks.
- **Recommendation:** A hybrid approach suits HPDA best.

Future Work

- Extend to multimodal content (images/videos).
- Explore hybrid approaches in AI content detection.
- Robustness testing against paraphrasing attacks.

Q & A

Thank You!

Questions?

References

Janelle Shane. Don’t use AI detectors for anything important. *Fortune*, 2023.

Debora Weber-Wulff et al. Testing of detection tools for AI-generated text. *Intl Journal for Educational Integrity*, 2023.

Geoffrey A. Fowler. We tested Turnitin’s ChatGPT-detector. *The Washington Post*, 2023.

Cynthia Condit. LibGuides: AI Legal Research Guide. *Univ. of Arizona Law Library*, 2025.

Kirchenbauer et al. A Watermark for Large Language Models. *arXiv:2301.10226*, 2023.

L. Cao. Watermarking for AI Content Detection. *arXiv:2504.03765*, 2025.

Y. Fu et al. Watermarking conditional text generation. *AAAI Conf.*, 2024.

Z. Jiang et al. Watermark-based Attribution of AI. *arXiv:2404.04254*, 2024.

H.U. Khan et al. Identifying AI content using DistilBERT. *Scientific Reports*, 2025.