Uncertainty Estimation Methods for Countering Attacks on Machine-Generated Text Detectors

Valeriy Levanov

Moscow Institute of Physics and Technology

Course: My first scientific paper
(Strijov's practice)

Expert: A. V. Grabovoy

Consultant: A. E. Voznyuk

2025

Goal of research

Machine-generated texts detection task

- Develop robust detectors for machine-generated text
- Counter adversarial attacks (homoglyphs, paraphrasing, noise injection)
- Achieve high accuracy with low computational costs

Key hypothesis

Uncertainty estimation methods can provide resilient detection without continuous retraining across attack types

Literature Review

Key Publications on Uncertainty Estimation

- **Polygraph**: Fadeeva A. et al. "Polygraph: Uncertainty-Aware Detection of LLM-Generated Text", ACL 2023
- **M4GT**: Wang Y. et al. "M4GT: Benchmark for Machine-Generated Text Detection", NAACL 2024
- **RAID**: Sadasivan V. et al. "RAID: Robust AI Detection Dataset", NeurIPS 2023

Recent Preprints

- **Image Uncertainty**: Jun Nie et al. "Detecting AI-Generated Images via Uncertainty", arXiv:2412.05897 (2024)
- **Perplexity Networks**: Pablo Miralles-González et al. "Token Weighting for AI Text Detection", arXiv:2501.03940 (2025)

Problem Statement

Binary text classification

Given:

- Input space $\mathcal{T}$ - all possible texts
- Output space $\mathcal{Y} = \{0, 1\}$ (0=human, 1=machine)

Detection model

Find mapping:

$$F : \mathcal{T} \to \{0, 1\}$$

that correctly classifies texts

Hypothesis

Machine-generated texts exhibit quantifiable differences in prediction confidence compared to human texts

Problem statement

Decomposed model architecture

$$F = f_3 \circ f_2 \circ f_1 : \mathcal{T} \to \{0, 1\}$$

where:

- $f_1 : \mathcal{T} \to \mathcal{L}$ - extracts context logits using LLM (Llama-3-8B)

$$f_1(t) = \{\ell_i\}_{i=1}^{L},\ \ell_i \in \mathbb{R}^{|V|}$$

- $f_2 : \mathcal{L} \to \mathbb{R}^d$ - computes uncertainty metrics
- $f_3 : \mathbb{R}^d \to \{0, 1\}$ - binary classifier

Quality metrics

- ROC-AUC (primary)
- Training time (primary)
- Accuracy

Solution

Perplexity

$$PPL = \exp\left(-\frac{1}{L}\sum_{l=1}^{L}\log P(w_l|w_{<l})\right)$$

- Information-based method

Mean Token Entropy

$$H = -\frac{1}{L}\sum_{i=1}^{L}\sum_{j} P(w_j|w_{<i})\log P(w_j|w_{<i})$$

- Information-based method

MC Entropy

$$H_S = -\frac{1}{K}\sum_{k=1}^{K}\log P(y^{(k)}|x)$$

- Information-based method

Mahalanobis distance

$$MD = \sqrt{(h-\mu)^T\Sigma^{-1}(h-\mu)}$$

- Dencity-based method
- Method fits a Gaussian centered at the training data centroid μ with an empirical covariance Σ matrix

Computational Experiment

Model Configuration

- **LLM**: Llama-3-8B-Instruct
- **Features**:
  - Top-512 context logits per token
  - Max token count - 512

Datasets

M4GT (arXiv)

- 12K machine / 6K human
- 6 generation models
- Clean data (no attacks)

RAID (Reddit)

- 15K machine / 15K human
- 12 generation models
- 11 attack types

Classification Models

Baseline Model

- **ROBERTa-Base** fine-tuned
- Trained for 1 epoch

Uncertainty-Based Classifiers

- **Logistic Regression**
  - Linear baseline
  - Fast training
- **Random Forest**
  - 300 trees
  - Max depth $= 10$

- **Neural Network**
  - Architecture:
    - 4 linear layers
    - BatchNorm + Dropout
  - Training:
    - Adam optimizer
    - BCE loss
    - 300 epochs

Results:

Model	ROC-AUC	Accuracy	Train Time (s)
BERT Classifier	0.9954	0.9942	1489
Neural Classifier + UE	0.7942	0.8183	208
Random Forest + UE	0.7831	0.8103	6.77
Logistic Regression + UE	0.7317	0.7744	0.013

Table: Performance comparison on arXiv data from M4GT

Model	ROC-AUC	Accuracy	Train Time (s)
BERT Classifier	0.9532	0.9538	2362
Neural Classifier + UE	0.8977	0.8987	378
Random Forest + UE	0.8987	0.8992	10.7
Logistic Regression + UE	0.7258	0.7271	0.035

Table: Performance comparison on Reddit data from RAID

Key findings:

- Accuracy reduction of BERT Classifier on attacked dataset
- 200x faster than BERT with 5.5% performance drop

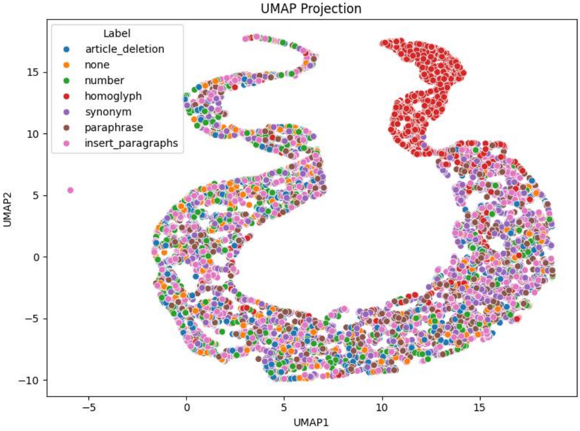
Conclusion

Results

- ROC-AUC: 0.89 (RAID dataset)
- Training time:
  - Rand Forest: 10s
  - Neural Net: 378s

Future Work

- Architecture search
- Hyperparameter optimization
- Attack pattern detection



UMAP: embeddings in uncertainty metric space by attacks

Uncertainty metrics reveal some attack patterns