EDMAL
ENHANCED DETECTION OF AI GENERATED TEXT

Introduction

The Problem

Recent developments in LLMs (Large Language Models, like ChatGPT) could:
- generate fake news
- create tensions between values
- be used for dishonest purposes

Current Solutions

Various AI-generated text detection methods:
- GPTZero, DetectGPT etc.
- **DNA-GPT** (Divergent N-Gram Analysis for Training-Free Detection of GPT-Generated Text)

Our Research

To find out if **Machine Learning** methods can enhance **DNA-GPT's** methods of text detection

Scope & Methodology

4 Methods

Stemming from the DNA-GPT's original methods, we propose 4 methods to enhance their text-detection:

Current DNA-GPT Detection using N-Gram Analysis

1. **Truncate** the text, *z*, using the truncate ratio of γ. (γ = 0.5)

"xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx" → z_2

2. **Regenerate** from the **truncated output, z_1,** using an LLM, K times. (K = 10)

$z_1 \rightarrow$ LLM $\rightarrow \{y_1, y_2, y_3, \vdots, y_k\}$ To be compared to z_2

3. **Compare** y_i to z_2 to classify *z* as AI-generated or Human-written
We propose and test **4 additional methods** of doing so along with doing analysis of n-grams (like DNA-GPT

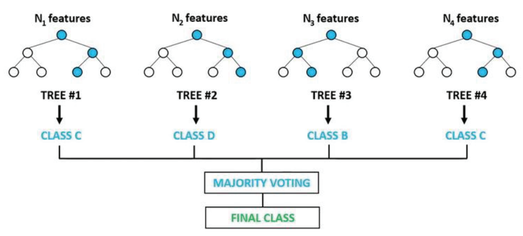
4. Using computed scores in step 3, and based on a **threshold**, evaluate if text is AI-generated or human-written
The threshold has to be fine-tuned to maximize scores for metrics

We grouped the 4 methods to improve step 3 into **2 groups:**

3.1 Extensions of N-gram Analysis

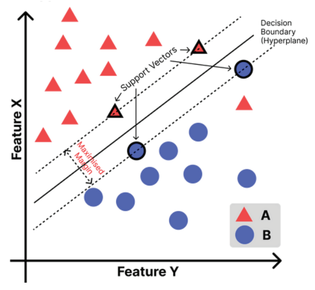
We trained both by taking feature vector *x* as input and returning a label of 1 (AI-generated) or 0 (Human-written) as *score_z*.

1 Random Forest Classifier



Extracted from an article on Medium titled: "Core Algorithms You Should Know in Classification" by Jwizzed.

2 SVM (Support Vector Machine)



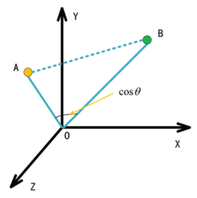
3.2 Alternative Approaches

3 Levenshtein Edit Distance (Does not use Machine Learning)

By taking the number of transformations (with **replace, insert and delete** operations) to change *y_i* to *z_2*, we aim to analyze the lexical structure of the text.

This method was added since its similar to the original N-gram analysis method DNA-GPT used.

4 Word Embeddings With Cosine Similarity



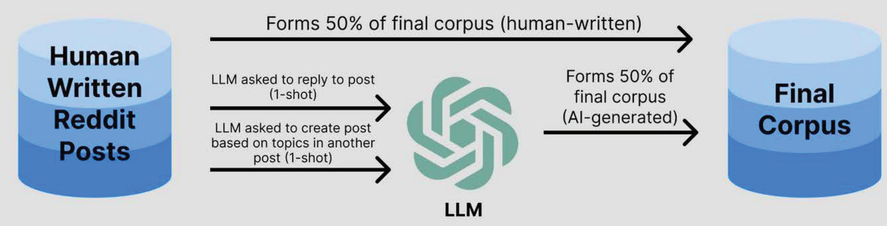
This method aims to capture the **semantic meaning** of text and **determine** texts' semantic distance.

Extracted from a paper from Semantic Scholar titled: "Cosine similarity based fingerprinting algorithm in WLAN indoor positioning against device diversity" by Shuai Han, Cong Zhao, W. Meng, Cheng Li

3 Datasets

2 datasets with differing minimum word counts were pruned and generated as follows:

Small Reddit Dataset (Min 100 & 500)

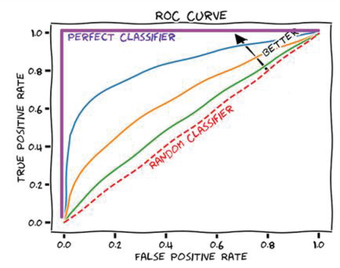


The final dataset, **ELI5 (Min 500),** was constructed as follows:
- **Human Section:** Replies to questions from the Explain Like I'm 5 community
- **AI Section:** Generated replies to the same questions

It was also used by DNA-GPT, allowing us to compare our results with DNA-GPT's.

2 Metrics

AUROC (Area Under Receiver Operating Characteristic Curve)



Extracted from an article on Glass Box Medicine titled: "Measuring Performance: AUC (AUROC)" by Dr Rachel Draelos

TPR (True Positive Rate) **at 1% FPR** (False Positive Rate)

This metric was used to ensure the **reliability** of detection algorithms for real-life deployment since it is crucial to maintain a **high TPR** while **minimizing the FPR**.

Both metrics were used by DNA-GPT, allowing for effective comparison of results.

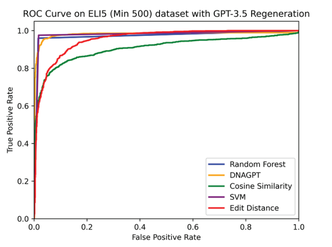
1 Model, Baseline and Detection Scenario

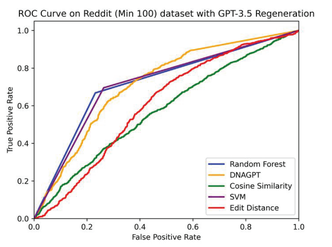
Due to time constraints, only one type of each were experimented with.
Model: GPT-3.5-Turbo, **Baseline:** DNA-GPT's original method, **Detection Scenario:** Black-Box

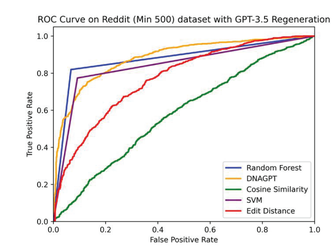
Results & Discussion

Scores for **AUROC** and **TPR** metrics for all **3 datasets** and **5 methods** (including DNA-GPT's original score), with the best-scores (with a margin of 1%) bolded red

Datasets	ELI5 (Min 500)		Reddit Small (Min 500)		Reddit Small (Min 100)	
Method	**AUROC**	**TPR**	**AUROC**	**TPR**	**AUROC**	**TPR**
DNA-GPT (original)	96.85	63.50	-	-	-	-
DNA-GPT	**98.07**	**59.08**	**88.32**	8.06	**71.50**	1.36
Random Forest	**97.20**	**61.04**	**87.60**	**12.16**	**71.83**	**2.89**
SVM	**97.91**	56.78	84.09	8.04	**71.62**	**2.64**
Cosine Similarity	90.75	38.41	58.07	1.11	57.58	1.13
Edit Distance	95.45	34.19	77.29	3.96	60.58	0.78







DNA-GPT
- Able to **closely match** their results with our attempt at replicating DNA-GPT
- Original N-Gram Analysis methods proved to be **extremely competitive**

Random Forest Classifier and SVM
- Random forest classifier consistently **performs the best** among the 5 methods
- SVM has inferior performance, possibly because Random forest classifier relies on **multiple models**
- However, both Random forest classifier and SVM require training, thus **with a larger dataset size**, improved results can definitely be achieved

Cosine similarity with Word Embeddings
- Performed **much worse** than other methods
- **Semantic meaning** of y_i is very likely to match z_2, with z_1 as context, irrelevant of whether z is human-written or AI-generated
- Scores **unable to properly differentiate** AI-generated from human-written

Edit Distance vs N-Gram Analysis
- Lexical analysis of edit distance is **much less effective** than n-gram analysis
- Edit distance is **unable to detect** words with similar spelling could have completely different meanings (eg. "Stationary" vs "Stationery")

Acknowledgements & References

[1] Yang, X. (2023, May 27). DNA-GPT: Divergent N-Gram Analysis for Training-Free Detection of GPT-Generated Text.
[2] Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. ELI5: Long form question answering. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 3558–3567
[3] Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. Detectgpt: Zero-shot machine-generated text detection using probability curvature. arXiv preprint arXiv:2301.11305, 2023.
[4] Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. arXiv preprint arXiv:2303.13408, 2023

Member:
Ang Jun Ray, Raffles Institution
Mentor:
Dr Chieu Hai Leong, DSO National Laboratories

YDSP Young Defence Scientists Programme
DSTA Defence Science & Technology Agency
DSO NATIONAL LABORATORIES