MANIPAL UNIVERSITY

INSPIRED BY LIFE

School of Computer Science and Engineering

Department of Computer Science and Engineering

Intrinsic Signals for Detecting AI-Generated Short Texts under Paraphrase Attacks (working title)

Submitted By:
Tanishq Choudhary
23FE10CSE00664

Supervised By:
Dr. Ashish Kumar

Outline

- Introduction
- Literature Review
- Problem Statement
- Proposed Solution
- Objectives

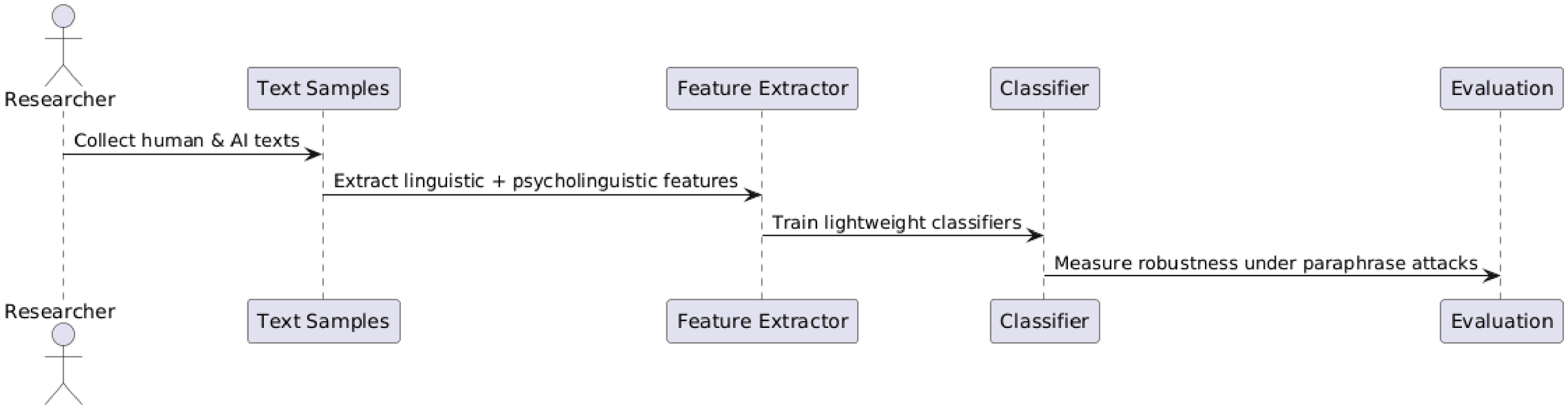
Introduction

- AI-generated text has become very common across education, software documentation, and communication.
- Traditional watermarking and classifier approaches fail when content is short (e.g., 50 to 200 tokens) or paraphrased (reworded by humans or other models). [1]
- There is a need for low-compute, interpretable, and robust detection approaches that can be reproduced on modest hardware.
- This work explores intrinsic, linguistic signals, stylometric, syntactic, and psycholinguistic to distinguish human vs. AI text even under rewriting.[2]

- References :

1. Opara, C. (2025). Distinguishing AI-Generated and Human-Written Text Through Psycholinguistic Analysis. arXiv preprint arXiv:2505.01800.
2. Mao, C., Vondrick, C., Wang, H., & Yang, J. (2024). Raidar: geneRative AI Detection via Rewriting. arXiv preprint arXiv:2401.12970.

Introduction



Literature Review

MGTBench: Benchmarking Machine-Generated Text Detection

- Presents a benchmark suite for detection across multiple LLMs and datasets.
- Emphasizes generalization to unseen models / domains as a core challenge.
- Highlights many detectors overfit to known generation distributions. VIEW SOURCE

MAGE: Machine-Generated Text Detection in the Wild

- Collects texts from diverse human sources + multiple LLMs.
- Shows the challenge of out-of-distribution detection: performance drops when encountering novel domains or unknown models. VIEW SOURCE

Literature Review

Stylometry recognizes human and LLM-generated texts in short samples (Przystalski et al., 2025)

- Uses stylometric features (lexical, syntactic, punctuation) to classify short texts (10 sentences).
- Reports high accuracy / Matthews correlation using decision trees / LightGBM and SHAP explanations. VIEW SOURCE

Opara, C. (2025): Distinguishing AI-Generated and Human-Written via Psycholinguistic Analysis

- Maps 31 stylometric features to cognitive/psycholinguistic dimensions (e.g. lexical retrieval, discourse planning).
- Shows interpretable alignment of features with cognitive theory, not just black-box models.
- VIEW SOURCE

Literature Review

Raidar: geneRative AI Detection via Rewriting (Mao et al., 2024)

- For a given text, ask an LLM to rewrite it; compute edit distance between original and rewrite.
- Hypothesis: AI-generated text will change less under rewriting compared to human text (because AI output is “already polished”).
- Shows improved F1 detection across domains (news, essays). VIEW SOURCE

Imitate Before Detect (Chen et al., 2024)

- Targets detecting machine-revised text: first learn a stylistic distribution model, then compare candidate texts’ deviation from that style.
- Demonstrates gains for detecting texts that are mixed (human + AI rewriting). Features with cognitive theory, not just black-box models. VIEW SOURCE

Problem Statement

Texts shortened to 50 to 200 tokens and paraphrased or rewritten remain under-detected by existing methods. There is little systematic work on robust stylometric/psycholinguistic features in short text + paraphrase settings.

Literature review gaps:

- Little focus on short texts (many studies use paragraphs or multi-sentence documents).
- Sparse analysis of which features survive paraphrase / rewrite attacks.
- Few works deliver CPU-friendly, interpretable pipelines (most rely on large neural detectors).

Objectives

- Design a feature set combining lexical, syntactic, psycholinguistic, and fluency proxy signals optimized for short texts.
- Evaluate detection models (LR, RF, etc.) under rewriting / paraphrase attacks (back-translation, LLM-based rewrite).
- Perform ablation / feature robustness study: which feature types degrade, and which remain stable.
- Produce a reproducible, low-compute framework suitable for MSc/SWE portfolios and reproducibility critiques.

Proposed Methodology

Data sources:

- Human text samples: HC3-English dataset (English human writing). (Yadagiri et al., 2024)
- AI-generated text: sample from GPT / open models.

Attacks / rewriting strategies:

- Back-translation (e.g. English to French to English)
- LLM-based rewriting: prompt GPT to “rewrite this text in different wording”
- Possibly edit distance / Raidar-style comparing rewrite vs original (as baseline)

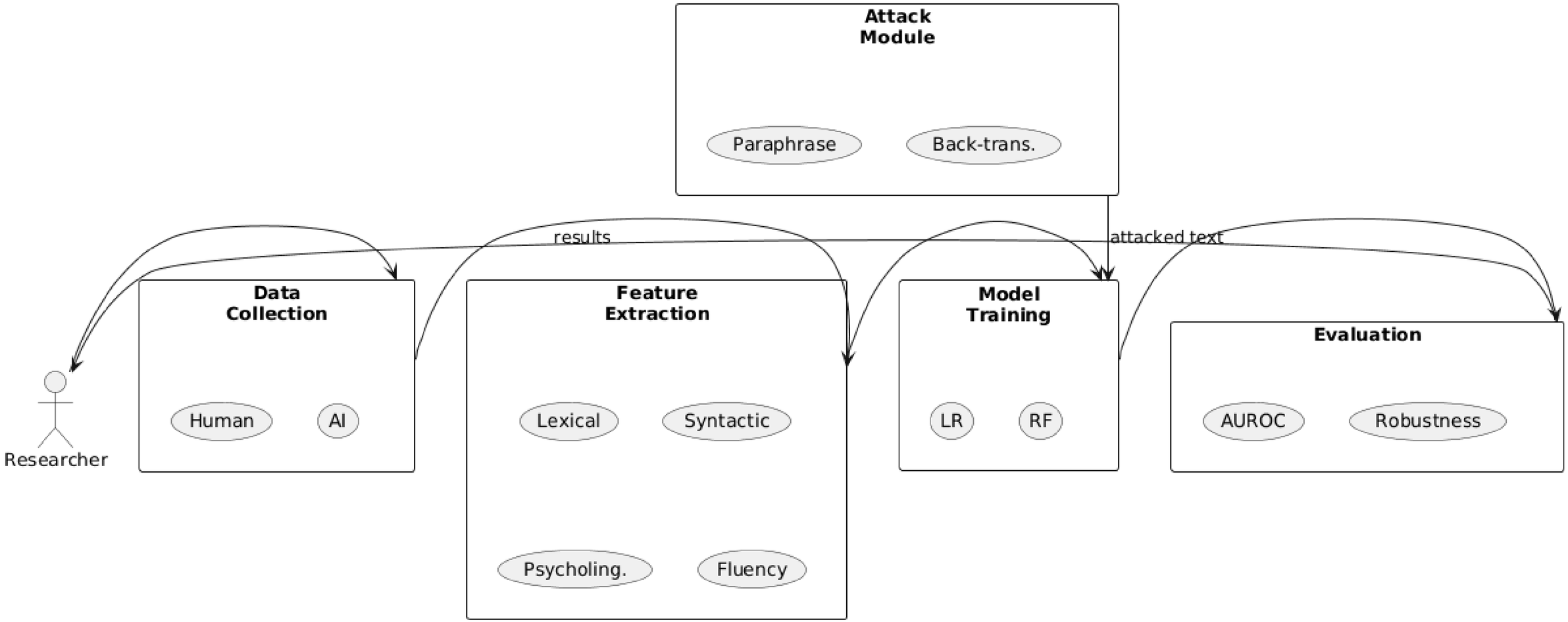
Feature extraction (interpretable, CPU-friendly):

- Lexical: type/token ratio, average word length, rare-word ratio
- Syntactic: POS tag distributions, parse tree depth (via spaCy)
- Psycholinguistic / cognitive: features from Opara’s set (e.g. mapping stylometric features to lexical retrieval, discourse planning)
- Fluency proxy: approximate log-prob from a small LM (e.g. GPT-2 or small transformer)

Models and evaluation:

- Train Logistic Regression / Random Forest / LightGBM
- Evaluate on original and attacked data, compute AUROC, drop deltas
- Perform ablation / feature importance (coefficients or SHAP)
- Compare performance with rewrite-based detectors (Raidar, Imitate Before Detect)

Proposed Methodology



Result

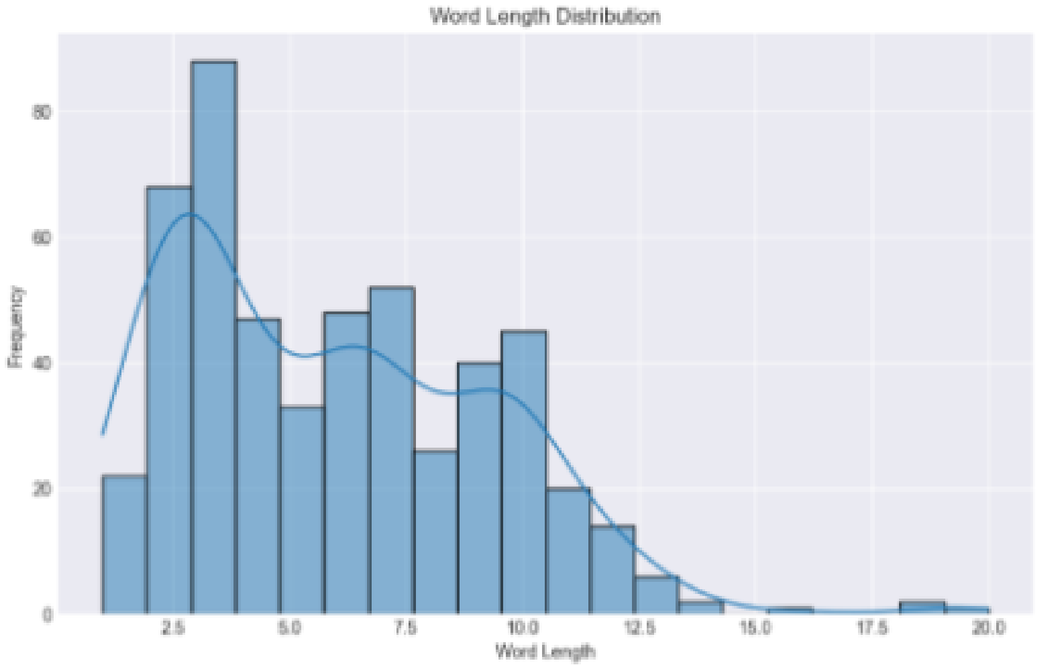
Text Analysis Report: theory_of_relativity_wikipedia

1. Basic Statistics

Metric	Value
Word Count	515
Sentence Count	26
Character Count	3633
Unique Word Count	270
Avg Word Length (chars)	5.85
Avg Sentence Length (words)	19.81
Avg Sentence Length (chars)	139.73
Sentence Length Std Dev (words)	12.18
Lexical Diversity (TTR)	0.5243
MATTR (Window=50)	0.7914

2. Entropy & Complexity

Metric	Value
Shannon Entropy (Words)	7.2313 bits
Shannon Entropy (Chars)	4.6032 bits
Function Word Ratio	0.3340



Character N-gram Analysis:

Common Character 3-grams (Top 10):

```
'the' (61), '_th' (49), 'ati' (43), 'al_' (37), 'lat' (33), 'he_' (32), '_re' (31), 'nd_' (31),
'vit' (30), 'and' (30)
```

Common Character 4-grams (Top 10):

```
'_the' (49), 'the_' (32), 'lati' (32), 'rela' (29), 'elat' (29), '_rel' (28), '_and' (28),
'and_' (28), 'ativ' (27), 'tivi' (25)
```

Common Character 5-grams (Top 10):

```
'_the_' (29), 'relat' (29), '_rela' (28), 'elati' (28), '_and_' (28), 'lativ' (27), 'ativi'
(25), 'tivit' (24), 'ivity' (23), 'vity_' (21)
```

Major details, see full folder for more

References

- He, X., Shen, X., Chen, Z., Backes, M., & Zhang, Y. (2023). MGTBench: Benchmarking Machine-Generated Text Detection.
- Mao, C., Vondrick, C., Wang, H., & Yang, J. (2024). RAIDAR: Generative AI Detection via Rewriting. Paper presented at the 12th International Conference on Learning Representations (ICLR 2024).
- Opara, C. (2025). Distinguishing AI-Generated and Human-Written Text Through Psycholinguistic Analysis. arXiv.
- Spiegel, M., & Macko, D. (2024). IMGTB: A Framework for Machine-Generated Text Detection. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL Demos).
- Wu, J., et al. (2025). A Survey on LLM-Generated Text Detection. Computational Linguistics, 51(1)
- Dugan, L., Hwang, A., Trhlik, F., Ludan, J. M., Zhu, A., Xu, H., Ippolito, D., & Callison-Burch, C. (2024). RAID: A Shared Benchmark for Robust Evaluation of Machine-Generated Text Detectors.
- Yang, X., et al. (2024). A Survey on Detection of LLMs-Generated Content. In Findings of EMNLP 2024.

Thank You