RUTGERS

Who wrote that?: Detecting the Source Model of Generated Text

Project Group 8

Aparajita Sarkar, Daniel Ojeda, Sritama Ghosh

Department of Computer Science

Rutgers University

Motivation

- LLMs are rapidly improving: Modern LLMs (GPT-5, GPT-4, Claude 4.5, Gemini) generate text that can be indistinguishable from human writing. This capability is beneficial, but creates a provenance problem.

- As AI models become more advanced and trained on broader data, their writing styles become harder to tell apart.

- This makes source identification increasingly challenging, even when we know the text is AI-generated If text is AI-generated, we often don't know which model produced it.

Key challenge:

Detection is no longer just "AI vs human."

We need attribution: **which** AI model wrote this? When text is AI-generated, we often cannot reliably identify which model produced it.

Why Source Attribution Matters?

Lack of source attribution reduces transparency, accountability, and trust in AI-assisted content.

- **Accountability & transparency**: identify model origin for audits and disclosure
- **Misinformation response**: trace model families behind coordinated content
- **Education & integrity**: understand tool usage without relying on self-reporting
- Enhances **AI safety and authenticity** efforts
- Supports accountability in content generation
- Enables better governance for model usage and provenance

Why Attribution Is Challenging?

- Multi-model attribution is significantly harder than binary AI detection
- Recent studies show early systems reach only about ~50% F1-score in multi-model attribution.
- Strong models produce outputs with overlapping style, making attribution unreliable.
- As models converge in quality and training data expands, their outputs become more similar,reducing separability.
- The problem is not just detecting AI text, it's distinguishing between highly capable models whose outputs can look very similar

Methodology Overview: Stylometric Feature Extraction

The main objective is to achieve high accuracy in attributing text to the correct source model based on Stylometric Features of the text.

- We aim for watermark-independent source attribution
- Leverage inherent writing patterns in model outputs (stylistic/statistical fingerprints)
- Systematically characterize cues that correlate with the originating model
- Our initial experiments focus on interpretable, stylometric features designed to capture surface-level writing patterns without relying on neural embeddings or black-box representations

At this stage, we study feature importance to understand what separates models.This stylometric feature evaluation is set as a baseline for model attribution.

These 59 features form a transparent baseline for attribution + feature-importance analysis.

Where Model Fingerprints Hide?

Stylometric Features	**What it captures**	**Example features (normalized where applicable)**
Lexical Features	verbosity, sentence construction, vocabulary usage, and punctuation habits	#chars/#words/#sentences; avg/min/max word length; avg/min/max sentence length; vocab size; type–token ratio; hapax ratio; punctuation density
Function Word Features	tokenization/alignment/decoding differences	Frequency of common function words (the, and, to, is, of, …)
Structural Features	how text is formatted and segmented, how models organize information, introduce explanations, separate ideas, and structure longer responses.	#paragraphs; avg paragraph length; newline ratio; quotes/parentheses/dashes/semicolons/colons
Statistical Distribution Features	how words are distributed	token entropy; unique-token ratio; repeated-type ratio; token frequency mean/max/SD
Local Discourse Coherence and Repetition Features	sentence-to-sentence lexical overlap	longest repeated-token run; repeated-token proportion

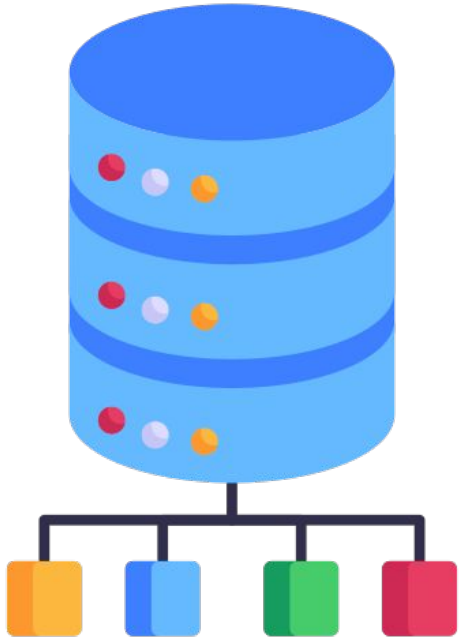
Limitations

- **No access to newest benchmark:** Could not evaluate on LLMtrace, limiting conclusions about generalization to newer model versions and updated generation behavior.
- **Hard-to-separate open models:** Most errors occur among LLaMA, Mistral, and MPT, likely due to overlapping training corpora suggesting possible intrinsic attribution ceiling for stylistically similar models.
- **Computing constraints:** Full hyperparameter sweeps (especially for transformer fine-tuning) were too expensive, could lead to missed higher-performing configurations.
- **Resource constraints:** Resource constraints restricted experiments to BERT/DistilBERT, preventing systematic testing of stronger/newer transformer backbones that might improve accuracy.

Dataset and Experimental Setup

- **Dataset:** RAID benchmark
- **Models:** Cohere, GPT-4, LLaMA, Mistral, MPT

- **Total texts:** 213,936
- **Split:** 80 / 20 stratified
  - **Train:** 171,149
  - **Test:** 42,787



Stylometric Baseline: Linear vs Tree-Based Models

Table 1: Classification performance using 59 stylometric features.

Classifier	Accuracy	Macro F1
Logistic Regression	0.587	0.563
Linear SVM	0.572	0.537
Random Forest	0.742	0.737
ExtraTrees	0.737	0.731
XGBoost	0.762	0.761
LightGBM	0.766	0.765

Table 2: Confusion matrix for LightGBM using 59 stylometric features.

	Cohere	GPT-4	LLaMA	Mistral	MPT
Cohere	3437	121	236	934	620
GPT-4	74	4432	472	258	112
LLaMA	108	232	9345	899	113
Mistral	616	232	1244	7241	1364
MPT	456	151	249	1518	8323

POS-Based Features

Table 4: Classification performance using stylometric and POS features (66 features).

Classifier	Accuracy	Macro F1
Random Forest	0.746	0.741
XGBoost	0.770	0.769
LightGBM	0.773	0.773

Table 5: Confusion matrix for LightGBM using stylometric and POS features.

	Cohere	GPT-4	LLaMA	Mistral	MPT
Cohere	3504	107	234	870	633
GPT-4	75	4456	444	251	122
LLaMA	88	225	9376	891	117
Mistral	600	218	1195	7342	1342
MPT	440	124	256	1464	8413

Character TF-IDF

Table 7: Classification performance using character-level TF-IDF features.

Classifier	Accuracy	Macro F1
XGBoost	0.776	0.776
LightGBM	0.793	0.795

Table 8: Confusion matrix for LightGBM using TF-IDF features.

	Cohere	GPT-4	LLaMA	Mistral	MPT
Cohere	3634	112	268	750	584
GPT-4	57	4716	349	139	87
LLaMA	108	137	9601	648	203
Mistral	519	130	958	7663	1427
MPT	478	94	361	1449	8315

Character TF-IDF and Stylometry Features

Table 10: Classification performance using TF-IDF and stylometric features.

Model	Accuracy	Macro F1
XGB_v1	0.8282	0.8313
XGB_v2	0.8345	0.8379
XGB_v3	0.8250	0.8279
XGB_v4	0.8135	0.8157
LGBM_v1	0.8485	0.8521
LGBM_v2	0.8528	0.8566
LGBM_v3	0.8586	0.8621
LGBM_v4	0.8558	0.8592

Table 12: Classification performance with extended stylometric feature set.

Model	Accuracy	Macro F1
Random Forest	0.7338	0.7235
XGBoost	0.8266	0.8293
LightGBM	0.8392	0.8423

Ensemble Methods: Voting and Stacking

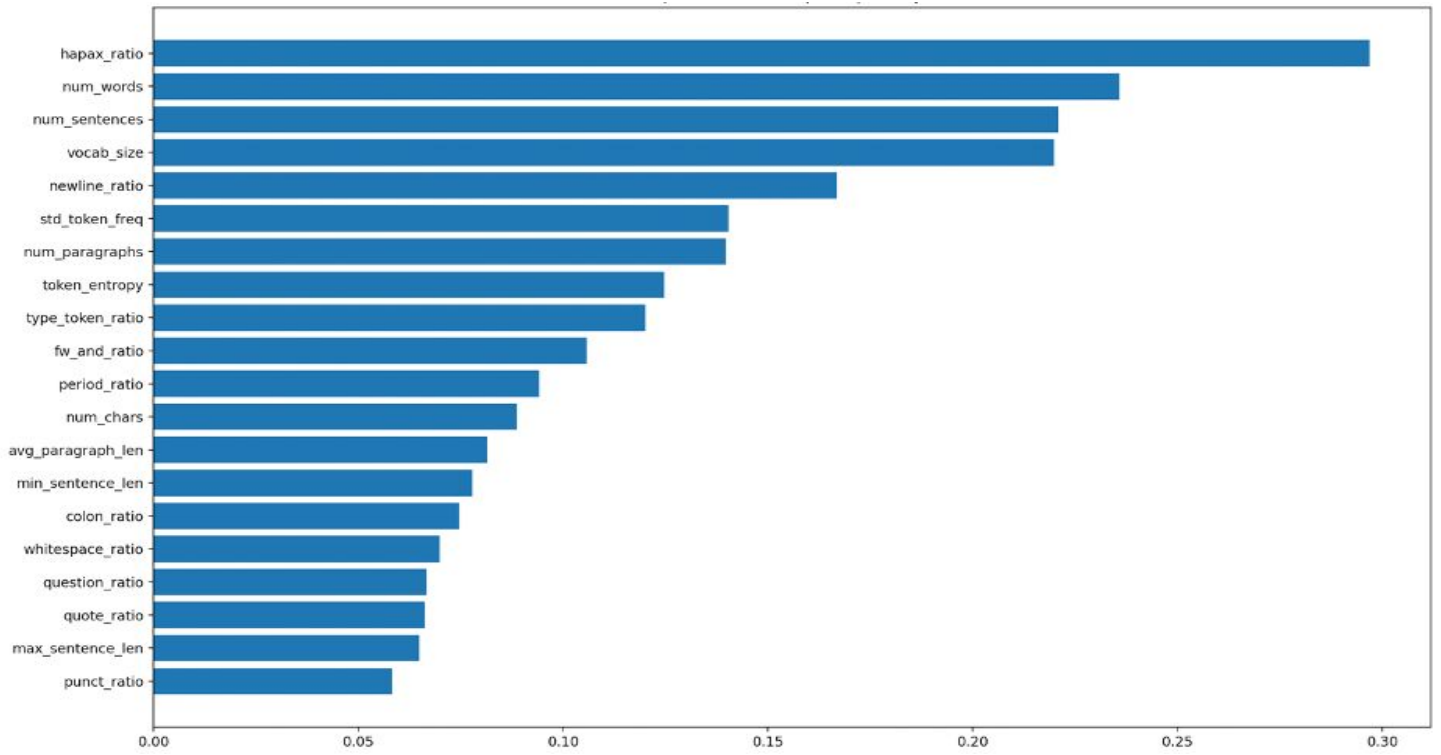
Table 13: Best voting ensemble performance.

Voting Ensemble	Base Models	Weights	Accuracy	Macro F1
VOTE-1	LGBM_v3 + XGB_v2	1:1	0.8525	0.8560
VOTE-2	LGBM_v3 + XGB_v2	2:1	0.8558	0.8595
VOTE-3	LGBM_v3 + XGB_v2 + LogReg	2:1:0.5	0.8545	0.8581
VOTE-4	LGBM_v3 + XGB_v2 + LogReg + SGD	2:1:0.5:0.5	0.3237	0.3003

Table 14: Stacking performance with meta-classifiers (base: LGBM_v3 + XGB_v2).

Meta-Classifier	Accuracy	Macro F1
XGBoost	0.8612	0.8649
ExtraTrees	0.8605	0.8644
Logistic Regression	0.8600	0.8638
SGD (log-loss)	0.8593	0.8631
LightGBM	0.8575	0.8615

Importance of Stylometric Features



The Way Forward - Transformer-Based Approach

	Model	Accuracy
Best ML model	LightGBM	86.18%
Best ML ensemble	LGBM + XGB + XGBoost meta classifier	86.12%
Goal	??	at least 90%

BERT v/s DistilBERT

Base Model	bert-base-uncased	distilbert-base-uncased
Parameters	110M	66M
Epochs	3	4
Accuracy	89.50%	89.78%
Macro F1	89.82%	90.17%

ML v/s Transformer

	Model	Accuracy
Best ML model	LightGBM	86.18%
Best ML ensemble	LGBM + XGB + XGBoost meta classifier	86.12%
Best DL model	DistilBERT	89.78%
Goal	??	at least 90%

Voting Ensemble models with DistilBERT

Method	Weights	Accuracy	Macro F1
Hard Voting	Equal	86.06%	86.54%
Soft Voting	Equal	90.08%	90.47%
Weighted Voting	By Accuracy	90.28%	90.67%
BERT Boosted	ML:BERT=1:2	91.17%	91.55%

Stacking Ensemble models with DistilBERT

Base Models	Meta-Features	Meta-Classifier	Accuracy	Macro F1
RF + XGB + LGB + DistilBERT	Probabilities (20-dim)	Logistic Regression	91.78%	92.18%
RF + XGB + LGB + DistilBERT	Predictions (4-dim)	Logistic Regression	85.92%	85.50%

Why is Stacking Better?

```
In the mystical land of Krynn, dragons rule the skies and shape the fate of its inhabitants.

As tensions rise between rival factions vying for con...

--------------------------------------------------------------------------
Ground Truth: mistral-chat
--------------------------------------------------------------------------
Model                       Prediction          Correct?
--------------------------------------------------------------------------
DistilBERT:                 mpt-chat            | N
LightGBM:                   llama-chat          | N
XGBoost:                    llama-chat          | N
Random Forest:              llama-chat          | N
Stacking Ensemble:          mistral-chat        | Y
```

Comparison of Ensemble with and without Random Forest

Method	With RF	Without RF	Winner
Stacking (LR)	0.9178	0.9174	With RF
BERT Boosted	0.9117	0.9100	With RF
Weighted Voting	0.9028	0.9080	Without RF
Soft Voting	0.9008	0.9069	Without RF
Hard Voting	0.8606	0.8704	Without RF

Final Comparison

	Model	Accuracy
Best ML model	LightGBM	86.18%
Best ML ensemble	LGBM + XGB + XGBoost meta classifier	86.12%
Best transformer model	DistilBERT	89.78%
Best ensemble with transformer	RF + XGB + LGB + DistilBERT	91.78%

THANK YOU