Detecting AI-Generated Text: An Interpretable Multi-Pipeline Approach Across Stylometric, TF-IDF, and Transformers Representations

Author: Aditi Jha; Advisor: Dr. John Musgrave
The College Of Wooster

Abstract

This study investigates how to distinguish human-written text from AI-generated text, addressing growing concerns around academic integrity, authorship, and plagiarism. Rather than focusing solely on classification accuracy, we examine the linguistic and representational patterns that differentiate human and AI writing across multiple modeling approaches. We evaluate four pipelines: stylometric features with classical machine learning, TF-IDF representations, fine-tuned transformer models, and a tokenizer-controlled TF-IDF setup to isolate vocabulary effects. Using a balanced dataset of 10,000 texts from diverse sources, the RoBERTa model achieved the highest accuracy (0.925), outperforming stylometric (0.77–0.85) and TF-IDF (0.76–0.80) approaches. Interpretability analyses (e.g., SHAP, LIME) and clustering revealed that traditional features capture broad trends but fail to clearly separate classes, while transformer embeddings show distinct semantic separation. These findings highlight that model architecture plays a larger role than feature representation, emphasizing the importance of interpretability for building reliable AI-text detection systems. Our study emphasizes the importance of understanding the representational signals behind predictions, underscoring the value of interpretability and evidence-based reasoning for developing more transparent detection systems.

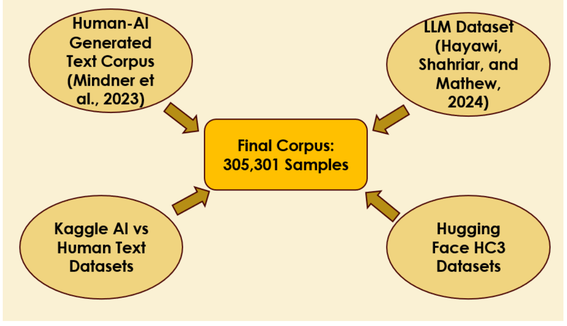
Research Questions

RQ1: How do different feature representation families: (a) handcrafted stylometric and statistical features, (b) lexical TF-IDF features, and (c) contextual transformer-based embeddings, differ in their ability to distinguish AI-generated text from human-written text across diverse domains and writing styles?

RQ2: When models operate on an identical tokenizer produced input space, how do classical machine learning classifiers compare to transformer architectures, and what does this reveal about the effect of model architecture independent of representation?

RQ3: What linguistic, stylistic, and semantic cues do different representations rely on during detection, and how do interpretability patterns differ across stylometric, lexical, and contextual pipelines?

Dataset



Methodology

Results

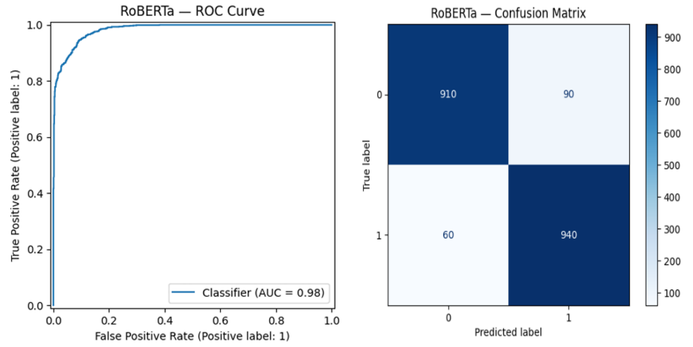
Summary statistics for perplexity features grouped by class (0 = Human, 1 = AI).

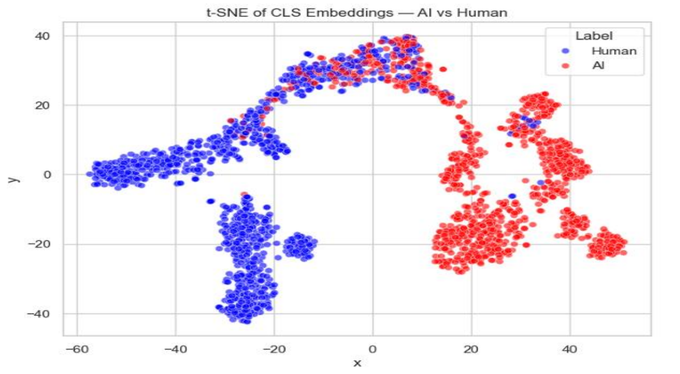
	Human (0)			AI (1)		
Feature	Mean	Median	Std	Mean	Median	Std
Mean Perplexity	187.733	98.685	530.310	66.627	38.301	389.120
Max Perplexity	378.107	141.239	1837.688	98.634	53.982	403.860

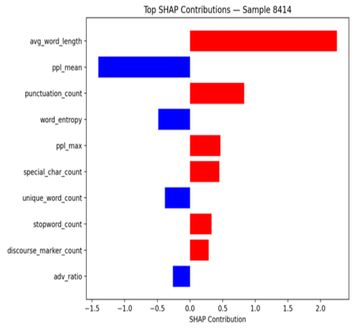
Performance Summary of All Pipelines

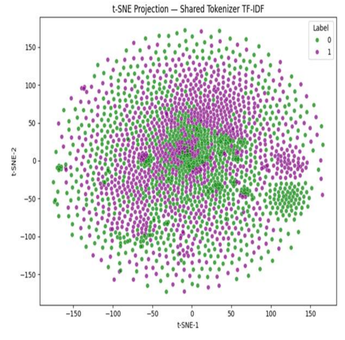
Pipeline	Model	Accuracy	Precision	Recall	F1 Score	ROC-AUC
Pipeline A: Stylometric Features	XGBoost	0.854	0.848	0.862	0.855	0.940
	Random Forest	0.842	0.846	0.835	0.840	0.933
	SVM (RBF)	0.820	0.813	0.833	0.823	0.908
	Logistic Regression	0.773	0.759	0.801	0.779	0.851
Pipeline B: TF-IDF	Logistic Regression	0.780	0.767	0.806	0.786	0.870
	Linear SVM	0.778	0.772	0.790	0.781	0.863
	RBF SVM	0.783	0.765	0.816	0.790	0.872
	Naïve Bayes	0.757	0.722	0.837	0.775	0.849
	Random Forest	0.806	0.854	0.738	0.792	0.878
	XGBoost	0.796	0.824	0.752	0.786	0.891
	MLP	0.766	0.757	0.785	0.771	0.871
Pipeline C: Transformer (RoBERTa)	Human (0)	0.925	0.938	0.910	0.924	0.98
	AI (1)	0925	0.913	0.940	0.926	0.98
Pipeline D: Shared Tokenizer	Logistic Regression	0.808	0.798	0.824	0.811	0.900
	Linear SVM	0.804	0.803	0.804	0.804	0.895
	RoBERTa	0.925	0.913	0.940	0.926	0.983

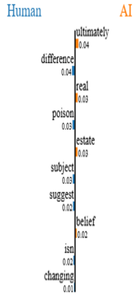
Transformers outperform classical methods by capturing deeper contextual signals in AI-generated text.











Text with highlighted words

I'd suggest changing the subject when your friends talk about real estate to save your sanity and friendship. There's a difference between "belief" and "knowledge". Arguing with a believer isn't a very productive course of action, and will ultimately poison the friendship. Reality is a harsh mistress.

Main Findings

- Transformer models achieve the highest performance by capturing contextual and semantic relationships, outperforming stylometric and TF-IDF approaches.
- Model architecture matters more than tokenization : even with shared tokenization, transformers outperform classical models.
- Stylometric and TF-IDF features capture useful but limited signals, reflecting surface-level patterns rather than deep structure.
- Surface-level patterns (e.g., n-grams) are insufficient: human and AI texts show highly similar lexical distribution
- AI-generated text is more probabilistically predictable, while human writing exhibits greater variability
- AI vs human text differs across multiple linguistic levels, but reliable detection requires deep contextual and probabilistic representations, not surface statistics.

Limitations

- Rapid evolution of LLMs may reduce detectable patterns over time (model drift).
- Evaluation performed on a combined dataset, limiting true cross-domain generalization.
- Binary classification does not capture hybrid human–AI authorship.
- Interpretability methods provide approximate (not exact) explanations of model behavior.

Future Work

- Evaluate models under cross-dataset and cross-domain settings.
- Develop methods for detecting hybrid and partially AI-generated text.
- Improve robustness to paraphrasing and adversarial editing.
- Extend detection to multilingual and emerging language models.

References

1. Hayawi, K., Shahriar, S., & Mathew, S. (2024). *The Imitation Game: Detecting Human and AI-Generated Texts.* Journal of Information Science..
2. Hua, H., & Yao, C. (2024*). Investigating Generative AI Models and Detection Techniques.* Frontiers in Artificial Intelligence.
3. Chaka, C. (2023). *Detecting AI content in responses generated by ChatGPT, YouChat, and Chatsonic.* Journal of Applied Learning and Teaching.
4. Abdali, S. et al. (2024). *Decoding the AI Pen: Techniques and Challenges in Detecting AI-Generated Text.* KDD.
5. Martinelli, F. et al. (2024). *A Method for AI-Generated Sentence Detection through Large Language Models.* Procedia Computer Science.

Acknowledgements

The author gratefully acknowledges the Mathematics and Computer Science (MCS) Department at The College of Wooster for providing access to NVIDIA A5000 GPU resources.