Hybrid Feature-Embedding Models for Robust AI Text Detection

Kasper Thomas Gartside Knudsen
IT-University of Copenhagen
Copenhagen, Denmark
kasper.knudsen@gmail.com

Christian Hardmeier
IT-University of Copenhagen
Copenhagen, Denmark
chrha@itu.dk

Abstract

Reliably detecting AI-generated text is crucial but challenged by AI evolution and evasion techniques. We introduce a set of linguistic features designed to discriminate between human and AI-generated text, and propose a hybrid detection model combining features with DistilBERT embeddings at the input level. This hybrid approach is evaluated against feature-only (XGBoost) and fine-tuned transformer (DistilBERT) methods. Furthermore, we contribute a challenging evasive test set, generated using an LLM unseen in the training data (Llama-3-8B), to evaluate performance across LLM domains and evasion techniques. While DistilBERT leads on the standard test set, our hybrid model demonstrates improved robustness against evasion, maintaining high precision on evasive texts. Hybrid feature-embedding models offer a promising approach towards building more resilient AI text detectors.

1 Introduction

Modern large language models (LLMs) generate human-like text, creating risks of misuse in academia, phishing, and disinformation, necessitating reliable AI text detection. Common classification methods include feature-based approaches, leveraging quantifiable linguistic features like term frequencies and syntactic patterns, and fine-tuned transformers, which capture deep contextual nuances. The rapid evolution of LLMs demands rigorous evaluation of detector robustness against out-of-distribution LLMs and evasive text generation.

Addressing this, we introduce novel linguistic features shown to improve the performance of feature-only classifiers in this domain. We also propose and evaluate a novel hybrid model concatenating our extensive collection of features with DistilBERT (Sanh et al., 2020) embeddings at the input level. The rationale is to combine the explicit pattern recognition of features with the deep semantic understanding of embeddings, potentially yielding a detector that is both high-performing and robust. We conduct a comparative analysis of our hybrid model against multiple baselines, including feature-only XGBoost (Chen and Guestrin, 2016) variations and a fine-tuned DistilBERT classifier. Evaluation focuses on performance against challenging texts generated by an LLM unseen during training (Llama-3-8B; Grattafiori et al., 2024), including outputs prompted to evade detection, to assess practical robustness.

2 Related Work

Research in AI text detection encompasses various approaches. Feature-based methods utilize linguistic cues, comparing strategies such as TF-IDF versus detailed handcrafted feature sets with classifiers like XGBoost (Schaaff et al., 2023; Shijaku and Canhasi, 2023). Concurrently, fine-tuned transformer models serve as powerful AI text detectors (Wang et al., 2024a). Relevant alternative techniques leverage AI model information differently, for instance through sentence-level analysis using LLM log probabilities (Wang et al., 2023) or via ensembles combining predictions from multiple detectors (Abburi et al., 2023; Zhang et al., 2024). A critical challenge across methods, however, is detector robustness against evasion; some studies develop adversarial generation techniques (Kumarage et al., 2023), while others note the limitations of current tools (Weber-Wulff et al., 2023). Our work focuses on document-level detection, investigating novel features, proposing an input-level hybrid feature-embedding model, and rigorously evaluating its robustness alongside established methods on challenging, evasive texts.

3 Methodology

This section details the dataset, the AI text detection classifiers implemented, and the methodology for evaluating robustness against evasive text.

Proceedings of the 21st Conference on Natural Language Processing (KONVENS), Volume 1: Long and Short Papers, pages 318–325, 2025

3.1 Dataset

We utilize the `ai_text_detection_pile` dataset from Hugging Face (artem9k, 2023), a large-scale collection of long-form human essays (from sources like Reddit, Q&A sites) and AI-generated texts generated by GPT models (GPT-2, GPT-3, GPT-J, ChatGPT). To ensure class balance for training and evaluation, we created a working subset of approximately 680,000 samples by sampling human texts to match the AI text count. Each sample includes the text (up to 1024 characters) and a binary *ai_generated* label.

3.2 Detection Approaches

We implement and compare three distinct detection strategies: feature-based models, a fine-tuned transformer model, and a novel hybrid model combining features and embeddings.

3.2.1 Feature-Based Detectors

We explored feature engineering using XGBoost classifiers trained on different feature sets. XGBoost models for each of these three feature configurations were independently optimized using randomized search with 3-fold cross-validation.

Baseline Features: We first established a comprehensive baseline of handcrafted features by replicating the work of Schaaff et al. (2023), excluding inapplicable dataset-specific features. This baseline includes 32 linguistic and statistical features, covering aspects like perplexity, semantics, document statistics, and readability, as well as text vector representations derived from TF-IDF (top 500 uni/bigrams) and Sentence-BERT (384-dim embeddings + average cosine distance). In total, this replicated baseline feature set comprises 917 features.

Extended Features: Our proposed contribution extends this baseline set by adding 8 new features. We hypothesized that AI text often exhibits an overly formal linguistic style that can feel distinct from typical human writing. To capture this *complexity*, we introduced these features: average dependent clauses, passive voice count, and average syntactic tree depth. Furthermore, observing AI's tendency to reuse specific words or phrases, we targeted *repetition* with the following features: n-gram entropy (uni, bi, and tri-grams), burstiness (indicates word clustering), and count of list items. A complete list of the extended features (which consists of the baseline features *and* our 8 proposed additions) is available in Appendix A, as well as mathematical definitions for select proposed features in Appendix B.

TF-IDF Model: As a much simpler baseline, we separately trained an XGBoost model using *only* the top 500 TF-IDF unigram and bigram features, replicating the work of Shijaku and Canhasi (2023).

3.2.2 Transformer-Based Detector

Representing a pure transformer approach, we fine-tuned DistilBERT (`distilbert-base-uncased`), a 66M parameter transformer, for binary text classification following guidelines from Hugging Face (2024). This approach leverages deep contextual understanding learned during pre-training. Input texts were lowercased and then tokenized to a fixed length of 512 tokens. Hyperparameters were optimized using Optuna (Akiba et al., 2019) over 10 trials based on validation performance, using mixed precision for efficiency.

3.2.3 Hybrid Feature-Embedding Detector

To investigate the synergy between linguistic patterns and contextual embeddings, we propose a hybrid model. This model combines our *extended* feature set, with sequence embeddings derived from a *pre-trained* (not fine-tuned) DistilBERT model. Specifically, we concatenate the feature vector with the embedding vector obtained via first token pooling (using the hidden state of the initial token as the sequence representation) from DistilBERT for each text sample. This combined vector feeds into a Feed-Forward Neural Network (FFNN) classifier, trained with the Adam optimizer (Kingma and Ba, 2015) and binary cross-entropy loss, and tuned using grid search.

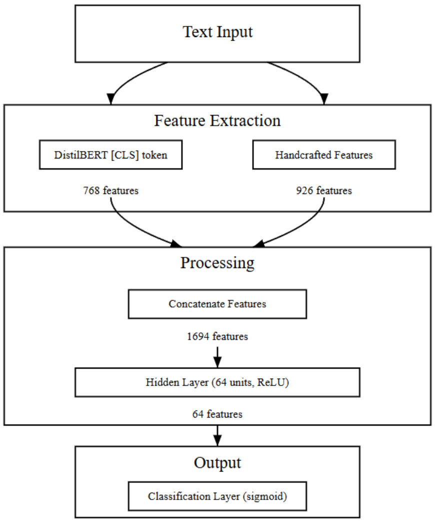

Figure 1: Hybrid FFNN Model Architecture

3.3 Evasive Text Generation

To assess practical robustness against out-of-distribution LLM behaviours and intentional evasion, we generated a challenging test set. We prompted an unseen LLM (Llama-3-8B) to rephrase 5,000 AI-generated texts from the original test set using three prompt strategies: *control* (rewrite in own style, a baseline for domain shift), *basic evasion* (rewrite to avoid detection), and *advanced evasion* (rewrite considering specific AI traits like repetition/formality to avoid detection) (see full prompts in Appendix C). Each set of 5,000 rephrased texts was combined with 5,000 human texts from the original test set, creating three balanced evasive evaluation datasets. All developed detectors were evaluated on these datasets to measure performance degradation and resilience.

4 Results

We evaluated the performance of the optimized detection models: feature-based XGBoost (baseline, extended, TF-IDF), fine-tuned DistilBERT, and our proposed hybrid FFNN. Performance was measured using accuracy, precision, recall, and F1-score on the standard test set (a 10% split of the dataset), and three evasive test sets generated using an unseen LLM (Llama-3-8B). Full result tables are available in Appendix D.

4.1 Performance on Standard Test Data

On the standard test set (Table 1), the fine-tuned DistilBERT model achieved the highest performance (98.3% F1). The baseline and extended feature-based XGBoost models also performed strongly; notably, the extended feature model achieved a small but consistent improvement of 0.1 percentage points over the baseline across all metrics (e.g., 96.6% vs. 96.5% F1). While this margin is slight, it suggests a positive contribution from our proposed linguistic features. Our hybrid FFNN model achieved a competitive 94.1% F1 score, while the TF-IDF XGBoost model lagged significantly (86.2% F1). This establishes strong performance baselines and supports the utility of feature engineering as a competitive detection strategy.

4.2 Robustness Against Evasive Texts

Evaluating models on the evasive datasets revealed varying degrees of robustness (Tables 2-4). While most detectors experienced severe performance degradation compared to the standard test set, the hybrid FFNN model demonstrated notable resilience. It exhibited comparatively slight F1 score degradation on the basic and advanced evasion texts and even improved performance on the control texts (Table 2). This observed variability across models and conditions underscores the necessity of evaluating detectors on diverse data, including outputs from unseen LLMs and text designed to evade detection, to assess practical resilience beyond potentially biased training distributions.

Regarding the feature-based detectors, the extended XGBoost consistently improved upon the baseline XGBoost across all evasive sets by at least 1% on F1-scores, further supporting the utility of the novel features against both domain shift and evasion attempts. However, both feature-based approaches showed vulnerability to the more targeted evasion strategies. Interestingly, the TF-IDF model substantially outperformed the baseline and extended XGBoost models on the basic and advanced evasion sets (e.g., 8.5% improvement on advanced evasion). This implies the handcrafted features were more susceptible to the prompt-based manipulations, while the TF-IDF representation proved more resilient, possibly due to the handcrafted set's complexity introducing noise or less robust patterns, indicating that more careful feature selection may be beneficial within such approaches.

Comparing the top-performing approaches on the evasive sets highlights the strengths of the hybrid FFNN relative to the fine-tuned DistilBERT. While DistilBERT mostly achieved higher F1 scores, the hybrid model demonstrated greater stability on evasive texts, evidenced by its significantly smaller F1 score degradation between the standard and advanced evasion test sets (a drop of only 4.9% vs. DistilBERT's 8.5%). Furthermore, the hybrid model outperformed DistilBERT on the control evasive set (F1: 94.6% vs 86.7%) and achieved F1 scores that were highly comparable on the basic and advanced evasion sets, 0.6-0.7% lower than DistilBERT. Critically, the hybrid model consistently maintained higher precision across all three evasive datasets, ranging from a 0.8% to 1.3% increase, indicating its positive classifications (AI) are more reliable against challenging and out-of-distribution text generation patterns.

4.3 Feature Importance Analysis

To interpret feature contributions, we applied SHAP (Lundberg and Lee, 2017) analysis to the

XGBoost extended feature model. Given its extensive feature space, it is significant that several of our novel handcrafted features ranked within the top 20 most impactful predictors globally (see Figure 2 in Appendix). Notably, *unigram entropy* ranked 3rd, *burstiness* ranked 11th, and *average dependent clauses* ranked 16th, with SHAP indicating higher values generally correlated with human predictions in this dataset. Their high importance confirms the contribution of these novel features to the detection model.

4.4 Impact of Evasion on Text Features

Analysis of mean values of features across labels and evasive texts (selected features shown in Table 5) reveals insights into model robustness challenges. Inherent differences existed between the training AI (GPT-family) and the unseen Llama-3 model (e.g., Llama-3 showed lower `burstiness` and `passive_voice_count`). Furthermore, evasive prompts manipulated features, reducing complexity (`dep_clauses_avg`) but sometimes yielded counter-intuitive results (e.g., reduced `error_count`, despite instructions to increase them). Averages for some novel features also appeared contradictory to model directionality learned via SHAP (Section 4.3). This discrepancy suggests that the model likely captures complex interactions or distributional nuances beyond simple averages. This combination of LLM differences and manipulated feature patterns explains the vulnerability of the feature-only XGBoost models to these challenging datasets.

5 Discussion

The performance degradation observed across all models when tested against our challenging evasive texts highlights the need for such datasets to provide a realistic assessment of resilience against evolving LLMs and adversarial tactics. The contribution of our extended features was most apparent on these challenging datasets, where they provided a clear improvement over the XGBoost baseline. Moreover, the hybrid model, which incorporates these same features, also exhibited notable robustness on these datasets. These findings demonstrate that feature engineering remains a valuable component in developing competitive and robust detection systems.

Our results also reveal a trade-off between performance on familiar data and robust generalization. While our fine-tuned transformer achieved the highest F1 scores on the standard test set, a finding consistent with similar benchmarks (Wang et al., 2024b), we argue its larger performance drop on evasive texts is the more critical result, suggesting potential sensitivity to distribution shift. In contrast, our novel hybrid FFNN model demonstrated superior resilience, evidenced by its minimal F1 degradation and consistently high precision on challenging evasive sets generated by an unseen LLM. This suggests combining linguistic features with contextual embeddings offers more stable detection against out-of-distribution patterns. The hybrid's high precision is particularly valuable for applications demanding low false positives. Although its peak performance didn't match DistilBERT here, its robustness profile suggests strong potential, possibly enhanced further via feature selection, embeddings from more powerful models, and expanded tuning.

However, even robust detectors like the hybrid model have inherent limitations and cannot serve as sole proof of misconduct (Weber-Wulff et al., 2023). We propose that practical systems, especially in academia, should integrate detection with verification techniques, for instance, analyzing a document's edit history via platform logs could reveal non-human generation patterns (e.g., large copy-pastes), though deliberate manual input might circumvent simple checks. Additionally, verifying reference validity could counteract LLM hallucination tendencies and outdated sources, providing a strong signal independent of writing style. Combining robust detectors with such verification checks, potentially programmatically, could create a more comprehensive system.

6 Conclusion

This research addressed the critical need for robust AI text detection by introducing linguistic features designed for this domain and proposing a novel hybrid model combining features with transformer embeddings. Through comparative analysis against strong feature-based (XGBoost) and fine-tuned transformer (DistilBERT) baselines, we evaluated performance on standard datasets and our challenging evasive datasets, which effectively exposed detector vulnerabilities. Our findings confirm that while fine-tuned transformers excel on in-distribution LLM data, the proposed hybrid model offers a compelling alternative, demonstrating greater robustness and stability against evasive texts while maintaining high precision. This performance is particularly advantageous for real-world scenarios requiring generalization to new text generation patterns and the minimization of costly misclassifications. Our proposed novel features also showed incremental benefits for feature-only models, particularly on evasive texts, confirming their value as discriminators of human and AI-generated text, and highlighting the necessity of robust evaluation against diverse data. Our results demonstrate that hybrid feature-embedding models serve as resilient and practically reliable AI text detection systems in the face of rapidly evolving AI capabilities.

7 Limitations

The generalizability of our findings is limited by our choice of data and models. Our training data consists of text from older AI models (GPT-2/3) and specific human genres (online essays). Similarly, our selection of DistilBERT for the fine-tuned and hybrid models constrained their potential performance. Employing larger, more contemporary transformers (such as ModernBERT (Warner et al., 2025)) was beyond our computational resources, but would likely yield stronger results and further widen the performance gap to non-transformer methods. Future work should incorporate more powerful models alongside datasets that are diverse and representative of the current AI landscape.

Furthermore, our experimental design limits the practical applicability of our findings. While this study focused on document-level detection, future work could extend our approach to incorporate sentence-level techniques for finer-grained analysis of potentially mixed human-AI texts, a common real-world scenario. Additionally, our use of a balanced dataset, while ensuring a fair comparison, does not reflect skewed distributions often found in practice. The relatively simple LLM rewriting prompts used for evasion also call for future work; more sophisticated evasive generation is needed for a thorough robustness assessment.

Finally, the proposed hybrid model has limitations regarding architecture and efficiency. The model has considerable inference overhead from dual feature extraction and embedding generation per input, contrasted with the potentially faster standalone fine-tuned model. Fast inference is particularly important for resource-constrained or real-time scenarios. Moreover, the hybrid's specific FFNN architecture, chosen feature set, and transformer embedding source represent just one possible implementation; exploration of alternative configurations may yield improved results.

8 Ethical Considerations

Our work aims to contribute positively to the challenge of identifying AI-generated text, with applications in supporting academic integrity and combating misinformation. However, we acknowledge the potential for misuse and the inherent risks of AI detection technology. As shown in our robustness evaluation, no detector is perfect, and performance can worsen on out-of-distribution or evasive text. Therefore, we argue that these tools should not be used as the sole basis for high-stakes decisions (e.g., determining academic misconduct). There is a significant risk of false positives, which could have serious negative consequences for individuals. We advocate for their use as an assistive tool for human-led evaluation, as discussed in Section 5. Finally, while the analysis of model weaknesses is necessary for improving detectors, we also recognize it could be used by adversaries to develop more effective evasion techniques.

9 Reproducibility

This study used the publicly available `ai_text_detection_pile` dataset (artem9k, 2023), which is compiled from various public online sources. To further support the reproducibility of our findings and to facilitate future research, all code and data created for this study are publicly available in a GitHub repository.[1] This includes the source code for all stages of our research, including feature extraction, model training, and evaluation, as well as the generated evasive text datasets.

The experiments were computationally intensive and conducted on high-performance computing (HPC) resources. Feature extraction and XGBoost model training were performed on multi-core CPUs (up to 64 cores), with system RAM requirements approaching 1000 GB for loading pre-computed embeddings. The fine-tuning of DistilBERT and the training of our hybrid FFNN model utilized NVIDIA A100 GPUs (40-80 GB VRAM).

[1] https://github.com/kakn/ai-text-detection

Acknowledgements

The authors acknowledge the IT University of Copenhagen HPC resources made available for conducting the research reported in this paper.

References

Harika Abburi, Kalyani Roy, Michael Suesserman, Nirmala Pudota, Balaji Veeramani, Edward Bowen, and Sanmitra Bhattacharya. 2023. A simple yet efficient ensemble approach for AI-generated text detection. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 413–421, Singapore. Association for Computational Linguistics.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A Next-generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, page 2623–2631, New York, NY, USA. Association for Computing Machinery.

artem9k. 2023. AI Text Detection Pile. Hugging Face Dataset. MIT License. Accessed: 2024-05-19.

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794, New York, NY, USA. Association for Computing Machinery.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 181 others. 2024. The Llama 3 Herd of Models. *arXiv e-prints*, arXiv:2407.21783.

Hugging Face. 2024. Sequence Classification Guide. Accessed: 2024-05-30.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Tharindu Kumarage, Paras Sheth, Raha Moraffah, Joshua Garland, and Huan Liu. 2023. How reliable are AI-generated-text detectors? an assessment framework using evasive soft prompts. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1337–1349, Singapore. Association for Computational Linguistics.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *Preprint*, arXiv:1910.01108.

Kristina Schaaff, Tim Schlippe, and Lorenz Mindner. 2023. Classification of human- and AI-generated texts for English, French, German, and Spanish. In *Proceedings of the 6th International Conference on Natural Language and Speech Processing (ICNLSP 2023)*, pages 1–10, Online. Association for Computational Linguistics.

Rexhep Shijaku and Ercan Canhasi. 2023. ChatGPT Generated Text Detection. Preprint available on ResearchGate.

Hao Wang, Jianwei Li, and Zhengyu Li. 2024a. Ai-generated text detection and classification based on bert deep learning algorithm. *Preprint*, arXiv:2405.16422.

Pengyu Wang, Linyang Li, Ke Ren, Botian Jiang, Dong Zhang, and Xipeng Qiu. 2023. SeqXGPT: Sentence-level AI-generated text detection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1144–1156, Singapore. Association for Computational Linguistics.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Osama Mohammed Afzal, Tarek Mahmoud, Giovanni Puccetti, and Thomas Arnold. 2024b. SemEval-2024 task 8: Multidomain, multimodel and multilingual machine-generated text detection. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2057–2079, Mexico City, Mexico. Association for Computational Linguistics.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.

Debora Weber-Wulff, Alla Anohina-Naumeca, Sonja Bjelobaba, Tomáš Foltýnek, Jean Guerrero-Dib, Olumide Popoola, Petr Šigut, and Lorna Waddington. 2023. Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, 19(1).

Ye Zhang, Qian Leng, Mengran Zhu, Rui Ding, Yue Wu, Jintong Song, and Yulu Gong. 2024. Enhancing Text Authenticity: A Novel Hybrid Approach for AI-Generated Text Detection. In *2024 IEEE 4th International Conference on Electronic Technology, Communication and Information (ICETCI)*, pages 433–438.

A Extended Collection of Features

Group	Feature
Sentiment	Polarity
	Subjectivity
Semantics	Stop Word Count
	Special Character Count
	Punctuation Count
	Quotation Count
	Uppercase Words Relative
	Personal Pronoun Count
	Personal Pronoun Relative
	POS Per Sentence Mean
	Discourse Marker Count
Structure	Character Count
	Word Count
	Sentence Count
	Paragraph Count
	Words Per Sentence Mean
	Words Per Sentence Std
	Words Per Paragraph Mean
	Words Per Paragraph Std
	Sentences Per Paragraph Mean
	Sentences Per Paragraph Std
Lexicon	Unique Word Count
	Unique Words Per Sentence Mean
	Unique Words Per Sentence Std
	Unique Words Relative
Readability	Flesch Reading Ease
	Flesch Kincaid Grade
Complexity*	Passive Voice Count
	Syntactic Tree Depth Per Sentence Mean
	Dependent Clauses Per Sentence Mean
Repetition*	Burstiness
	Unigram Entropy
	Bigram Entropy
	Trigram Entropy
	List Item Count
Errors	Error Count
	Multi Blank Count
Text Vectors	500-dim TF-IDF Vector (Uni/Bigram)
	Sentence-BERT Vector
	Average Sentence-BERT Distance
LLM-based	Max Perplexity
	Mean Perplexity
	Zero-Shot LLM Prediction

*Groups containing our proposed features

B Formulations of Select Features

This section provides mathematical details for proposed features whose implementation is not self-evident from their names:

- **Average Syntactic Tree Depth**: The average depth of syntactic parse trees across all sentences in a text.

  $$\text{ASTD} = \frac{1}{N}\sum_{i=1}^{N} d_i$$

  Where N is the number of sentences and d_i is the depth of the syntactic tree for the i-th sentence.

- **N-gram Entropy**: Measures the entropy of n-gram distributions, with higher entropy indicating more variety.

  $$H(X) = -\sum_{x\in X} p(x)\log p(x)$$

  Where X is the set of n-grams and $p(x)$ is the probability of n-gram x, calculated from its frequency.

- **Burstiness**: Measures the variability in word frequencies, with higher values indicating more word clustering.

  $$B = \frac{\sigma_w}{\mu_w}$$

  Where σ_w is the standard deviation and μ_w is the mean of word frequencies.

C Evasive Text Generation Prompts

To generate the evasive text datasets, we prompted Llama-3-8B (Grattafiori et al., 2024) to rewrite AI-generated texts using the following instructions. An instruction to "Respond with only the rewritten text." followed each prompt request:

- **Control Prompt** (AI-rephrased): "Rewrite the following text in your own style and tone."
- **Basic Evasion Prompt**: "Rewrite the following text so it won't be detected by AI detection tools."
- **Advanced Evasion Prompt**: "Considering typical AI text traits like repetitive phrasing, overly formal language, and minimal errors, rewrite the following text so it won't be detected by AI detection tools."

D Classification Results

Model	Accuracy	Precision	Recall	F1
DistilBERT (Fine-tuned)	98.3%	97.0%	99.7%	98.3%
FFNN (Hybrid)	94.1%	92.8%	95.4%	94.1%
XGBoost (Baseline)	96.4%	95.6%	97.3%	96.5%
XGBoost (Extended)	96.5%	95.7%	97.4%	96.6%
XGBoost (TF-IDF)	85.7%	83.0%	89.7%	86.2%

Table 1: Performance metrics on 34,000 AI + 34,000 human texts

Model	Accuracy	Precision	Recall	F1
DistilBERT (Fine-tuned)	87.8%	95.9%	79.1%	86.7%
FFNN (Hybrid)	92.1%	97.2%	92.1%	94.6%
XGBoost (Baseline)	71.5%	63.9%	98.8%	77.6%
XGBoost (Extended)	73.2%	56.4%	98.7%	78.7%
XGBoost (TF-IDF)	72.3%	77.2%	63.2%	69.5%

Table 2: Performance metrics on 5000 AI-rephrased + 5000 human texts

Model	Accuracy	Precision	Recall	F1
DistilBERT (Fine-tuned)	91.5%	96.2%	86.4%	91.0%
FFNN (Hybrid)	86.5%	97.0%	84.5%	90.3%
XGBoost (Baseline)	68.1%	61.4%	97.6%	75.4%
XGBoost (Extended)	69.8%	62.7%	97.8%	76.4%
XGBoost (TF-IDF)	81.4%	81.3%	81.5%	81.4%

Table 3: Performance metrics on 5000 basic evasive AI + 5000 human texts

Model	Accuracy	Precision	Recall	F1
DistilBERT (Fine-tuned)	90.4%	96.1%	84.2%	89.8%
FFNN (Hybrid)	85.0%	96.9%	82.6%	89.2%
XGBoost (Baseline)	57.7%	54.2%	98.5%	70.0%
XGBoost (Extended)	60.3%	55.8%	98.5%	71.3%
XGBoost (TF-IDF)	80.0%	80.8%	78.8%	79.8%

Table 4: Performance metrics on 5000 advanced evasive AI + 5000 human texts

E SHAP Visualization

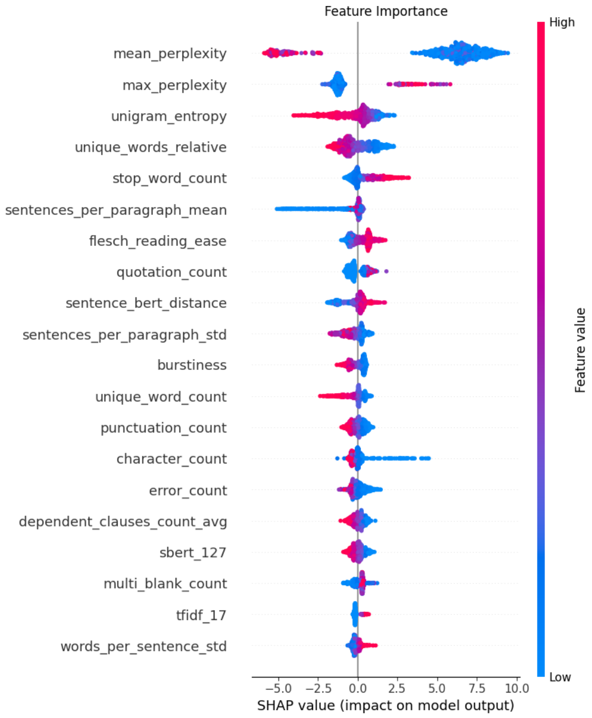

Figure 2: SHAP summary plot for the extended feature model

F Feature Means Across Labels

Feature	Human	AI	Control	Basic	Advanced
quotation_count	1.85	6.85	2.08	2.71	2.70
list_item_count	0.22	0.38	0.41	0.57	0.53
passive_voice_count	1.47	2.41	1.37	1.54	1.47
words_per_par_std	31.34	21.41	10.29	12.00	11.46
stop_word_count	128.72	184.00	104.24	102.23	99.17
discourse_count	22.70	32.30	20.63	18.85	19.27
dep_clauses_avg	1.17	1.65	1.53	1.34	1.33
error_count	8.40	6.01	1.93	2.28	2.45
burstiness	3.57	4.67	2.90	3.00	2.93
uniq_words_per_sent	17.08	21.87	21.01	20.44	20.45
syntactic_depth_avg	5.53	6.95	6.89	6.93	6.86
flesch_reading_ease	73.57	62.77	61.22	49.20	50.17
unigram_entropy	6.52	6.40	6.38	6.20	6.31

Table 5: Mean values of select features for human and AI texts (340k texts each) and various evasive texts (5k texts each, only AI label).