UDC 004.89:519.76
DOI https://doi.org/10.32782/tnv-tech.2025.2.15

STYLOMETRIC CLASSIFICATION OF AI-GENERATED TEXTS: COMPARATIVE EVALUATION OF MACHINE LEARNING MODELS

Petryshak T. V. *– Postgraduate Student*
Lviv Polytechnic National University
ORCID ID: 0009-0006-6296-3867

Rybchak Z. L. *– Candidate of Technical Sciences (PhD), Associate Professor,*
Associate Professor at the Department of Information Systems and Networks
Lviv Polytechnic National University
ORCID ID: 0000-0002-5986-4618

The rapid proliferation of large language models (LLMs), such as ChatGPT and Deepseek, has made it increasingly difficult to distinguish between AI-generated and human-written text. This study evaluates the effectiveness of stylometric analysis as a transparent and interpretable method for detecting synthetic content. A balanced dataset of 30,000 short-form responses (10,000 per class: Human, ChatGPT, Deepseek) was constructed. While the Human and ChatGPT responses were sourced from an existing dataset, the Deepseek responses were generated using standardized prompts to ensure consistency. Each response was transformed into a vector of 12 anually engineered features capturing lexical richness, syntactic structure, and readability. The study involved five classifiers: Logistic Regression, Support Vector Machine, Random Forest, Gradient Boosting, and Decision Tree. Each was trained and evaluated on multiclass and binary classification tasks. Randomized hyperparameter tuning was applied to enhance performance. The tuned Random Forest achieved the highest results, with macro-averaged F1-scores of 0.84 (multiclass) and 0.86 (binary), and accuracy over 87 %. Gradient Boosting and SVM showed comparably strong performance, confirming the robustness of ensemble and margin-based methods in this context. Key features such as Simpson's Index, type-token ratio, and sentence length proved most informative. The results confirm that stylometric features, despite their simplicity, can reliably distinguish between human and AI-generated text. The results indicate that this approach demonstrates clear potential and, when used in combination with other methods, can contribute effectively to the identification of AI-generated content. Additionally, generating datasets using open-source models with the Ollama framework enables affordable and scalable experimentation without relying on commercial APIs. This is particularly beneficial for early stage research and academic environments with limited resources.

Key words: *AI-generated text, stylometry, text classification, machine learning, large language models.*

Петришак Т. В., Рибчак З. Л. Стилометрична класифікація штучно згенерованих текстів: порівняльне оцінювання моделей машинного навчання

Із поширенням великих мовних моделей, таких як ChatGPT і Deepseek, дедалі складніше визначити, хто є автором тексту – людина чи штучний інтелект. У цьому дослідженні оцінюється ефективність стилометричного аналізу як прозорого та інтерпретованого методу для виявлення синтетичного контенту. Було сформовано збалансований набір із 30 000 відповідей (по 10 000 для кожного класу: Human, ChatGPT, Deepseek). Відповіді для Human і ChatGPT взято з відкритого датасету, а для Deepseek створено окремо за єдиним шаблоном запитів з використанням моделі Deepseek 7B. Кожну відповідь перетворено на вектор із 12 стилометричних ознак, що характеризують лексику, синтаксис та читабельність. Дослідження охоплює п'ять моделей машинного навчання: Logistic Regression, SVM, Random Forest, Gradient Boosting та Decision Tree. Кожну з них навчено та протестовано для багатокласової та бінарної класифікації з подальшою оптимізацією гіперпараметрів. Найвищу ефективність показала модель Random Forest (F1 = 0.84/0.86), досягнувши точності понад 87 %. Gradient Boosting і SVM також продемонстрували хороші результати. Найінформативнішими ознаками виявились індекс Сімпсона, співвідношення типів і токенів та середня довжина речень. Результати підтверджують,

що стилометричні ознаки, попри свою простоту, дозволяють ефективно розрізняти тексти людського та штучного походження. Запропонований підхід демонструє високу інтерпретованість і може бути ефективно використаний у поєднанні з іншими методами для верифікації авторства, забезпечення академічної доброчесності та виявлення згенерованого контенту. Крім того, генерація даних за допомогою відкритих локальних моделей у середовищі Ollama забезпечує масштабованість експериментів використання платних API, що особливо актуально на ранніх етапах досліджень та в академічному середовищі з обмеженими ресурсами.

Ключові слова: *штучно згенерований текст, стилометрія, класифікація текстів, машинне навчання, великі мовні моделі.*

Introduction. Large language models (LLMs) such as ChatGPT, Claude, and Deepseek are reshaping how people generate and interact with text across domains. These systems can produce fluent, contextually appropriate writing that closely resembles human output, making it increasingly challenging to determine whether a given text was written by a person or generated by AI. This issue has become particularly relevant in areas where authorship and authenticity matter most, including education, journalism, academic publishing, and content governance.

A number of detection tools have been developed in response to this challenge. Systems like GPTZero and DetectGPT often rely on token-level statistics or probability-based heuristics to assess textual origin. However, many of these tools are closed-source, commercial in nature, and lack transparency regarding their internal logic or limitations. Their accuracy can vary across different domains and languages, and their use may involve high computational or financial costs, especially for large-scale applications.

Stylometric analysis provides a complementary perspective grounded in linguistic theory. It focuses on measurable properties of writing style such as lexical diversity, sentence structure and readability which help characterize authorship. Although stylometry has a long history in traditional authorship attribution, it remains relevant today and continues to offer valuable insights when applied to texts generated by the latest generation of LLMs. This study evaluates how effectively interpretable machine learning models can distinguish between human- and AI-generated text using stylometric features. A balanced dataset of 30,000 labeled short-form answers was created, evenly distributed among human-written, ChatGPT, and Deepseek responses. Several classifiers were trained and evaluated in binary and multiclass settings using optimized parameters. The results show that even simple, manually engineered linguistic features can provide strong predictive power in identifying AI-generated text. Stylometric methods retain their value in the modern landscape, particularly when integrated into broader detection frameworks aimed at increasing the transparency and accountability of AI-generated content.

Objectives of the Study. The primary objective of this study is to evaluate the effectiveness of interpretable machine learning models in distinguishing between human-written and AI-generated text based on stylometric features. Additionally, this work demonstrates that local deployment of language models using the Ollama framework can serve as a practical and cost-effective strategy for generating labeled data and conducting controlled experiments in the domain of AI-generated text detection.

Review of Literature. Various methods have been proposed to detect AI-generated text, ranging from token-based classifiers to statistical and linguistic approaches. Tools like GLTR highlight improbable word choices based on language model probabilities [1], while models such as DetectGPT rely on curvature of log-probability functions to identify synthetic content [2]. Although these methods show promise, real-world detectors like GPTZero and Originality.AI exhibit inconsistent accuracy across domains, ranging from 55 % to 97 % [3], and are often closed-source and non-transparent.

Stylometric analysis offers a linguistically grounded alternative that focuses on quantifiable aspects of writing style, such as lexical richness, syntactic structure, and readability. Traditionally used in authorship attribution, stylometric features are now applied to AI detection. For instance, StyloAI used 31 manually engineered features to classify texts with up to 98 % accuracy on educational datasets, significantly outperforming GPTZero on paraphrased content [4]. Other studies combined stylometry with perplexity and semantic embeddings, achieving F1 scores exceeding 96 % [5].

Unlike black-box neural models, classical machine learning algorithms – such as logistic regression, SVM, decision trees, and ensemble methods – allow interpretability by exposing which features influence predictions [6]. Experiments show that Random Forest and Gradient Boosting, in particular, perform well when paired with robust stylistic features [4]. Key indicators like average sentence length, function word ratios, and pronoun usage have proven effective for differentiating between human and AI writing.

Despite these advances, current detectors face several challenges. Deep learning-based systems often act as "black boxes", making decisions difficult to interpret, which is problematic in education or journalism where transparency matters [5]. Moreover, minor input changes (e.g., abbreviating a word) can lead to misclassification [7]. Many models also lack robustness across content types and domains [4], highlighting the need for reliable and generalizable solutions.

A less-explored area is interpretable multi-class attribution: identifying not just whether text was AI-generated, but which model (e.g., ChatGPT vs Deepseek) produced it. The AuTexTification shared task emphasized this need [8], but most submissions used deep ensembles that sacrificed interpretability. Some early work reframes this as an authorship attribution task using stylometry [6], and hybrid approaches show potential [9], yet a transparent stylometry-based framework for fine-grained LLM attribution remains largely underdeveloped.

Dataset and Data Collection. The dataset used in this study was carefully curated to support both binary and multiclass classification tasks (Human vs AI, as well as Human vs ChatGPT vs Deepseek). It contains 30,000 labeled English-language responses evenly distributed across three classes: Human, ChatGPT, and Deepseek. To ensure consistency across model comparisons, a unified question set was used. Specifically, 10,000 unique question prompts were randomly sampled from the publicly available HC3 dataset [10]. For each question, the following responses were collected:

- *Human Response*: One corresponding answer manually written by a human from the HC3 dataset.
- *ChatGPT Response*: One matching answer generated by ChatGPT, also retrieved from HC3.
- *Deepseek Response*: A new answer generated using the Deepseek 7B model via the Ollama framework. The generation was automated using a Python pipeline with the following prompt structure: "Answer the following question: {question}". A schematic of the data generation pipeline is presented in Fig. 1.

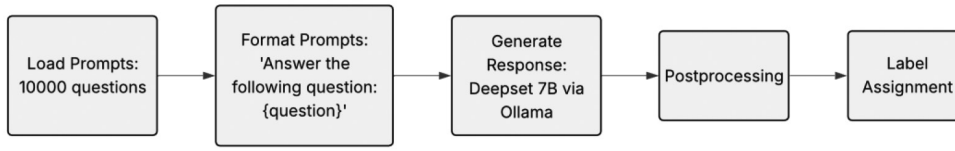

Fig. 1. Deepseek Response Generation Pipeline

The pipeline iterated over each question and recorded the Deepseek model's output. All responses were normalized and stored alongside the associated question and a corresponding label: «human», «chatgpt», or «deepseek». The resulting dataset contained exactly 10,000 responses per class, each aligned to the same prompt.

Responses were converted to lowercase, stripped of special characters, and cleaned of extra whitespace. No stemming or stopword removal was applied to preserve stylistic characteristics. Questions were retained in the dataset for context but were excluded from the classification feature extraction. All 30,000 responses are written in English. To ensure uniqueness, duplicate entries were removed using hash-based filtering. This balanced and structurally consistent dataset (Table 1) provides a solid foundation for assessing stylistic differences among human and LLM-generated texts.

Table 1

Class Distribution and Data Sources

Class	Number of Samples	Source
Human	10,000	HC3 Dataset [10]
ChatGPT	10,000	HC3 Dataset [10]
Deepseek	10,000	Generated via Ollama (Deepseek 7B)

Methodology. Each response in the dataset was transformed into a 12-dimensional feature vector representing its stylistic properties. The original answer text was removed, retaining only numerical representations of style. This preprocessing ensured consistency and interpretability across all downstream machine learning models.

The extracted features capture various aspects of writing style and structure, including vocabulary use, grammatical composition, and textual clarity. Table 2 provides a concise overview of all stylometric features used for classification.

To compute readability metrics, standard formulas were applied. The Flesch Reading Ease score is defined as:

$$FRE = 206.835 - 1.015 \times \left(\frac{totalwords}{totalsentences} \right) - 84.6 \times \left(\frac{totalsyllables}{totalwords} \right). \quad (1)$$

The Gunning Fog Index is calculated as:

$$Fog = 0.4 \times \left[\left(\frac{totalwords}{totalsentences} \right) + 100 \times \left(\frac{complexwords}{totalwords} \right) \right], \quad (2)$$

where "complexwords" are defined as those with more than three syllables.

After feature extraction, the dataset of 30,000 samples was stratified and split into training (80 %) and testing (20 %) subsets to maintain class balance. Feature values were normalized using z-score scaling:

$$x' = (x - \mu)/\delta, \quad (3)$$

x – raw feature value, μ – mean, and δ – standard deviation calculated from the training data. Standardization was especially important for models such as Logistic Regression and SVM, which are sensitive to feature scale.

In addition to the original three-class dataset, a binary version was constructed by merging ChatGPT and Deepseek into a single "AI" class, resulting in a balanced two-class dataset (Human vs AI). Five classical machine learning models were implemented

Table 2

Stylometric features extracted from each text sample

Category	Feature	Description
Lexical	Average Word Length	The mean number of characters per word, reflecting the vocabulary complexity. Texts with longer average word length may indicate more sophisticated vocabulary.
	Type-Token Ratio (TTR)	Ratio of unique words to total words; measures lexical diversity
	Simpson's Diversity Index	Captures repetitiveness in vocabulary; higher values indicate lower diversity
	Yule's K Index	A robust lexical diversity metric that accounts for word frequency distribution. It is essentially the probability that two randomly picked words from the text are the same. A lower diversity in an AI-generated text could signal repetitive word usage or over-reliance on common words, whereas human authors might introduce more unique or context-specific terms
Syntactic	Average Sentence Length	Average number of words per sentence; reflects syntactic complexity
	Punctuation Frequency	Rate of punctuation per word; captures stylistic variation
	Pronoun Ratio	Proportion of pronouns; reflects personal or impersonal tone
	Noun Ratio	Proportion of nouns; high values may indicate factual or list-like structure
	Verb Ratio	Proportion of verbs; shows degree of action-oriented language
	Function Word Ratio	Share of function words in the text; relates to grammatical structure
Readability	Flesch Reading Ease (FRE)	Indicates text simplicity; higher scores mean easier readability.
	Gunning Fog Index	Estimates years of education needed to understand the text.

using the scikit-learn framework: Logistic Regression, Support Vector Machine (SVM), Decision Tree, Random Forest, and Gradient Boosting. Each classifier was trained and evaluated independently on both the multiclass and binary versions of the dataset.

Hyperparameter optimization was performed using RandomizedSearchCV with cross-validation. Model performance was evaluated using Accuracy, Macro F1-score, and class-specific F1-scores.

Text preprocessing and feature extraction relied on several widely used Python libraries. spaCy was used for tokenization, sentence segmentation, and POS tagging. TextStat was used to compute syllable counts and readability metrics. NLTK provided stopword filtering and additional linguistic utilities. Data manipulation was carried out using pandas and NumPy.

The complete training pipeline for each task included: (1) feature standardization, (2) stratified splitting, (3) model fitting, and (4) evaluation. All models were implemented using the scikit-learn framework. Figure 2 illustrates the full experimental pipeline from raw feature matrix construction to final evaluation.

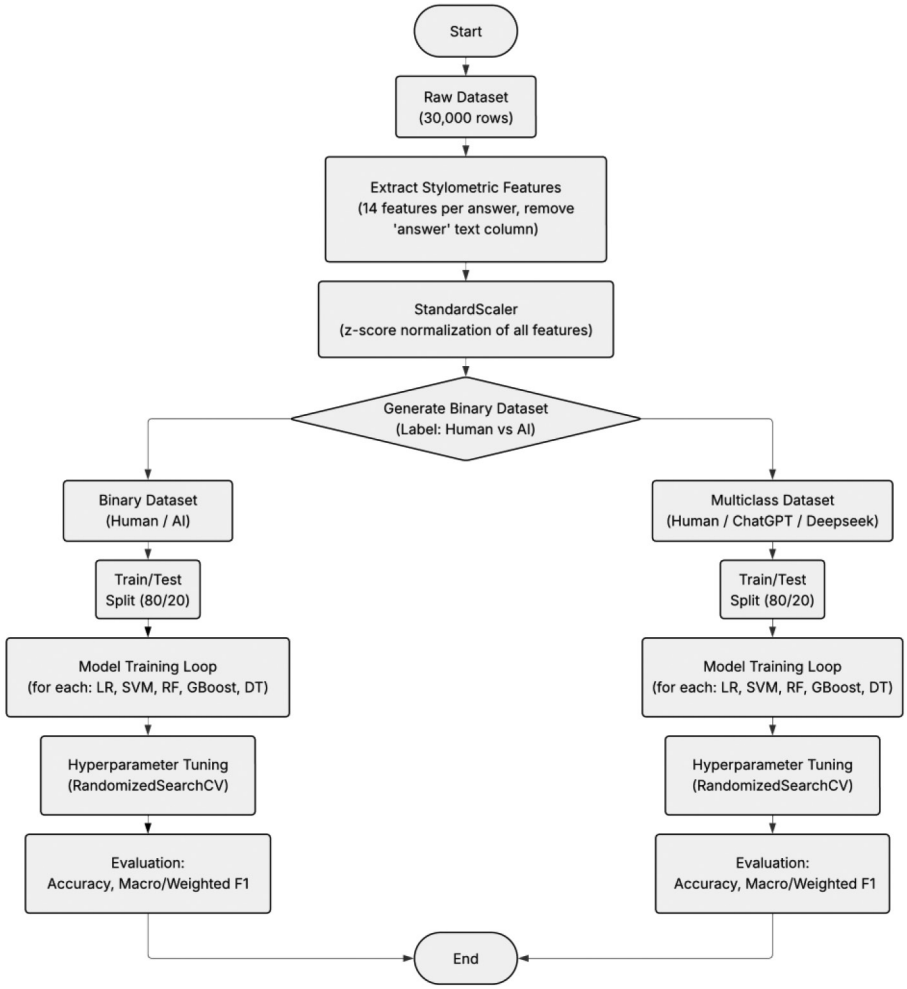

Fig. 2. Stylometric Classification Pipeline

Results and Discussion. The classification experiments were conducted in both multiclass and binary settings, using default and tuned hyperparameters for all models. The evaluation began with multiclass classification to distinguish among responses generated by humans, ChatGPT, and Deepseek.

Table 3 presents the classification metrics using default hyperparameters. Among all models, Random Forest and SVM consistently achieved the best results. Random Forest reached 84 % accuracy and a macro F1-score of 0.84, followed closely by SVM and Gradient Boosting. Logistic Regression and Decision Tree showed moderate performance, with the latter being the lowest-performing under default parameters.

Hyperparameter tuning improved classification metrics across most models, especially ensemble methods. As shown in Table 4, Gradient Boosting achieved the highest accuracy (84 %) and macro F1-score, demonstrating that optimized configurations enhance performance significantly. Random Forest and SVM also showed better class-wise balance. Logistic Regression improved only slightly, as it is

Table 3

Multiclass Classification Results (Default)

Model	Accuracy	F1 (ChatGPT)	F1 (Deepseek)	F1 (Human)	Macro F1
Logistic Regression	0.70	0.80	0.69	0.61	0.70
Random Forest	0.84	0.87	0.84	0.80	0.84
SVM	0.82	0.87	0.82	0.77	0.82
Decision Tree	0.75	0.81	0.76	0.70	0.75
Gradient Boosting	0.81	0.86	0.82	0.76	0.81

Table 4

Multiclass Classification Results (Tuned)

Model	Accuracy	F1 (ChatGPT)	F1 (Deepseek)	F1 (Human)	Macro F1
Logistic Regression	0.71	0.80	0.69	0.61	0.70
Random Forest	0.837	0.87	0.85	0.80	0.84
SVM	0.827	0.87	0.83	0.78	0.83
Decision Tree	0.79	0.83	0.81	0.73	0.79
Gradient Boosting	0.84	0.87	0.82	0.80	0.84

constrained by its linear nature. Decision Tree saw better generalization due to depth restriction and pruning.

Fig. 3 shows the confusion matrix for the tuned Gradient Boosting classifier. Most classification errors occurred between the ChatGPT and DeepSeek classes, indicating stylistic overlap between AI models. Human-authored text was more consistently identified, highlighting stronger distinctiveness in human writing.

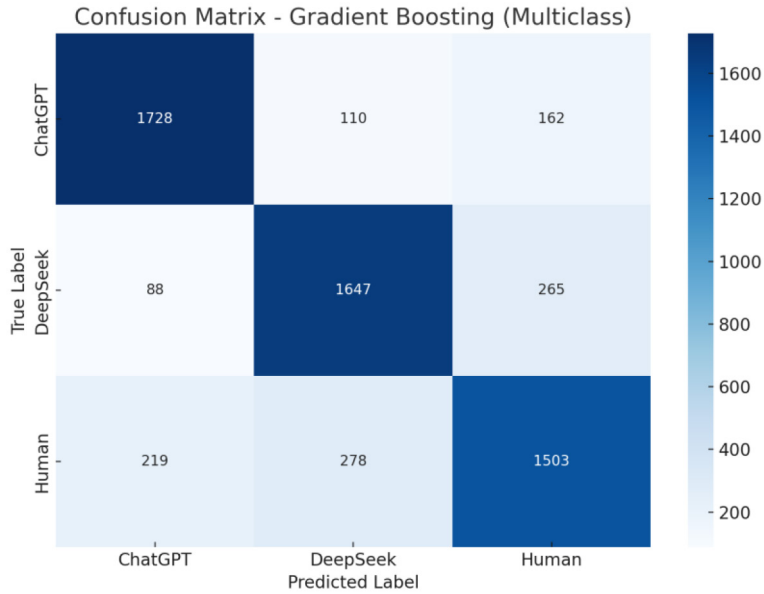

Fig. 3. Confusion matrix for the tuned Gradient Boosting classifier

Fig. 4 presents the confusion matrix for the tuned Random Forest classifier on the same multiclass task. Similar to Gradient Boosting, most errors occur between the two AI-generated categories, while human text remains the most accurately identified class. Random Forest exhibited slightly more misclassification of Deepseek as ChatGPT than Gradient Boosting, but overall class-wise balance was similar.

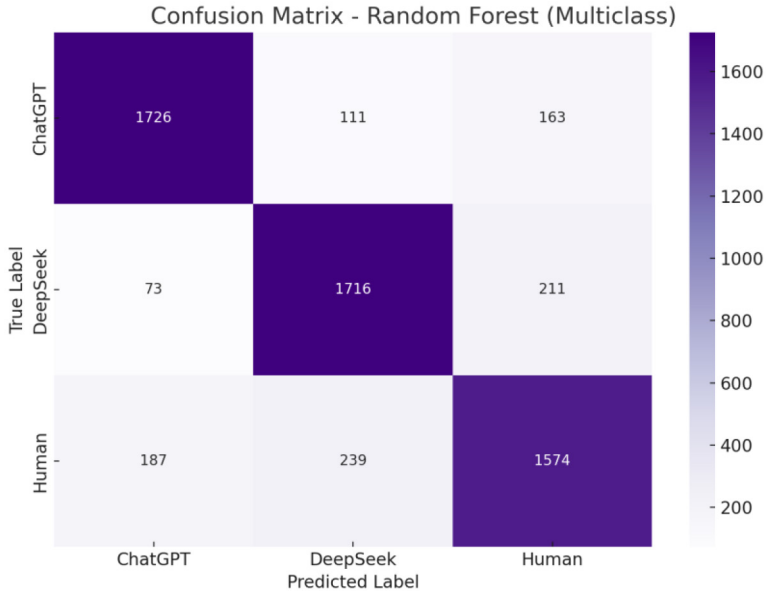

Fig. 4. Confusion matrix for the tuned Random Forest classifier

The binary classification scenario simplified the task to distinguishing human from AI-generated text. Under default settings, all classifiers improved notably. As shown in Table 5, ensemble models once again achieved the highest scores. Random Forest and Gradient Boosting surpassed 83 % accuracy. Even Logistic Regression reached 76.6 %, confirming the strength of stylometric features in separating the two classes.

Table 5

Binary Classification results (Default)

Model	Accuracy	F1 (AI)	F1 (Human)	Macro F1
Logistic Regression	0.766	0.84	0.58	0.71
Random Forest	0.866	0.90	0.79	0.84
SVM	0.834	0.88	0.72	0.80
Decision Tree	0.796	0.85	0.70	0.77
Gradient Boosting	0.838	0.88	0.73	0.81

Tuned models, shown in Table 6, further increased performance. Random Forest achieved nearly 88 % accuracy and a macro F1-score above 0.86. Gradient Boosting also crossed 87 %, confirming the value of hyperparameter tuning. Simpler models like Logistic Regression and SVM approached ensemble performance, indicating that the human vs AI distinction is linearly separable to a large extent.

Table 6

Binary Classification results (Tuned)

Model	Accuracy	F1 (AI)	F1 (Human)	Macro F1
Logistic Regression	0.778	0.84	0.61	0.73
Random Forest	0.879	0.91	0.81	0.86
SVM	0.869	0.90	0.79	0.85
Decision Tree	0.85	0.89	0.76	0.83
Gradient Boosting	0.87	0.91	0.81	0.86

Fig. 5 presents the confusion matrix for the tuned Random Forest in binary classification. The classifier exhibited high precision and recall, with most misclassifications being false negatives (AI misclassified as Human). This conservatism is desirable in real-world applications like authorship verification.

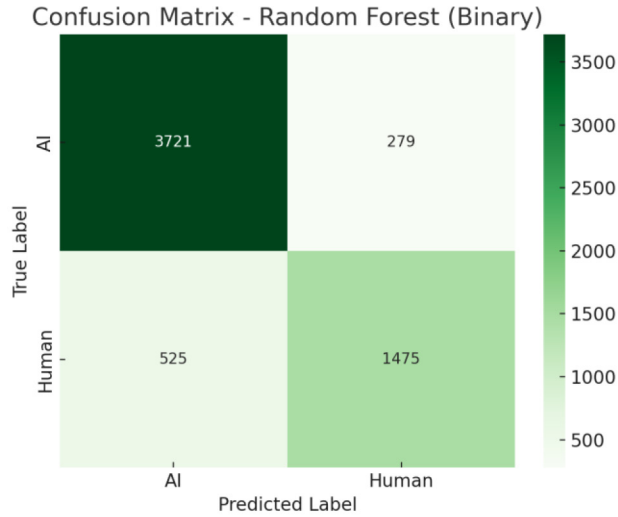

Fig. 5. Confusion matrix for the tuned Random Forest classifier

Fig. 6 shows the corresponding confusion matrix for Gradient Boosting. The pattern of classification errors closely resembles that of Random Forest, with both classifiers achieving high recall for human-written texts and only marginal confusion involving high-quality AI-generated content. The balanced distribution of predictions confirms the robustness of ensemble models in binary classification.

Table 7 summarizes the best-performing configurations for each model. These tuned settings were selected using randomized search and cross-validation and proved essential in boosting overall classification scores.

Fig. 7 compares macro F1-scores for all models before and after tuning, highlighting performance improvements due to optimized hyperparameters.

To better understand which stylometric features contributed most to model predictions, feature importance scores were extracted from the Random Forest and Gradient Boosting classifiers. Table 8 and Table 9 list the top-ranked features for each model. Across both models, Simpson's Diversity Index and type-token ratio consistently emerged as the most informative indicators. These metrics effectively capture lexical

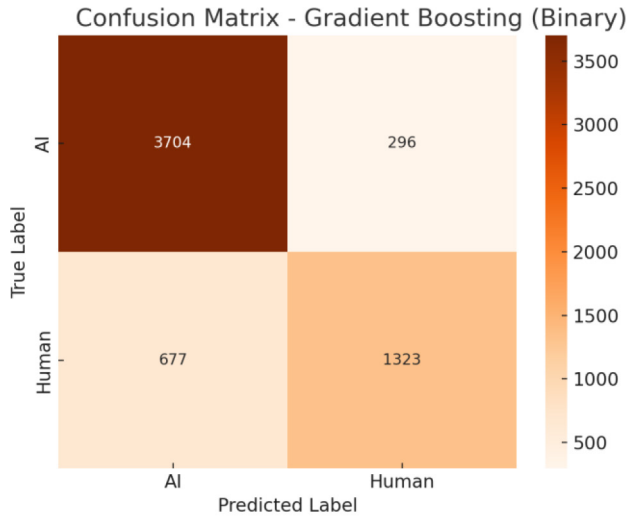

Fig. 6. Confusion matrix for the tuned Gradient Boosting

Table 7

Tuned Parameters

Model	Multiclass	Binary
Logistic Regression	'solver': 'saga', 'penalty': 'l2', 'max_iter': 3000, 'C': 0.1	'solver': 'lbfgs', 'penalty': 'l2', 'max_iter': 3000, 'C': 0.7
Random Forest	'n_estimators': 300, 'min_samples_split': 4, 'min_samples_leaf': 1, 'max_depth': 25	'n_estimators': 300, 'min_samples_split': 4, 'min_samples_leaf': 1, 'max_depth': None
SVM	'kernel': 'rbf', 'gamma': 'scale', 'C': 10.0	'kernel': 'rbf', 'gamma': 'scale', 'C': 100.0
Decision Tree	'min_samples_split': 2, 'min_samples_leaf': 1, 'max_depth': 10	'min_samples_split': 2, 'min_samples_leaf': 1, 'max_depth': 10
Gradient Boosting	'min_samples_split': 2, 'min_samples_leaf': 1, 'max_depth': 10	'n_estimators': 300, 'max_depth': 5, 'learning_rate': 0.1

variation and repetitiveness, making them reliable indicators for distinguishing between AI and human writing.

The consistent top-ranking of simpsons_d and type_token_ratio suggests that lexical diversity and richness are among the strongest indicators distinguishing AI from human writing.

Fig. 8 displays a SHAP summary plot for the tuned Gradient Boosting model, confirming that the top features identified by global importance measures also have the highest local impact on model predictions.

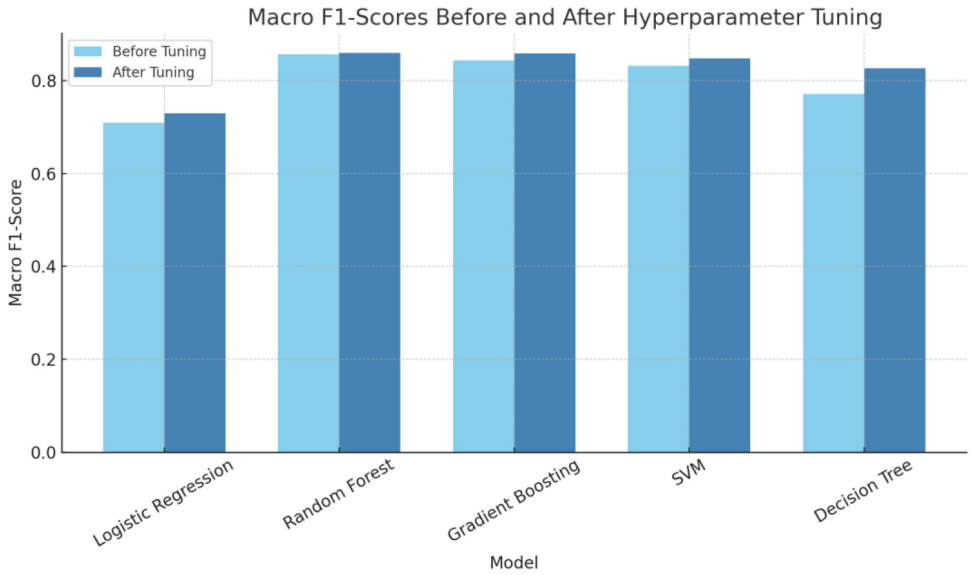

Fig. 7. Hyperparameter tuning results

Table 8

Feature Importances – Gradient Boosting

Feature	Importance
simpsons_d	0.309
type_token_ratio	0.191
avg_sentence_length	0.096
punctuation_ratio	0.089
avg_word_length	0.074
yules_k	0.069
gunning_fog_index	0.037
function_word_ratio	0.033

Table 9

Feature Importances – Random Forest

Feature	Importance
simpsons_d	0.184
type_token_ratio	0.142
avg_sentence_length	0.100
punctuation_ratio	0.093
yules_k	0.085
avg_word_length	0.070
flesch_reading_ease	0.065
gunning_fog_index	0.059

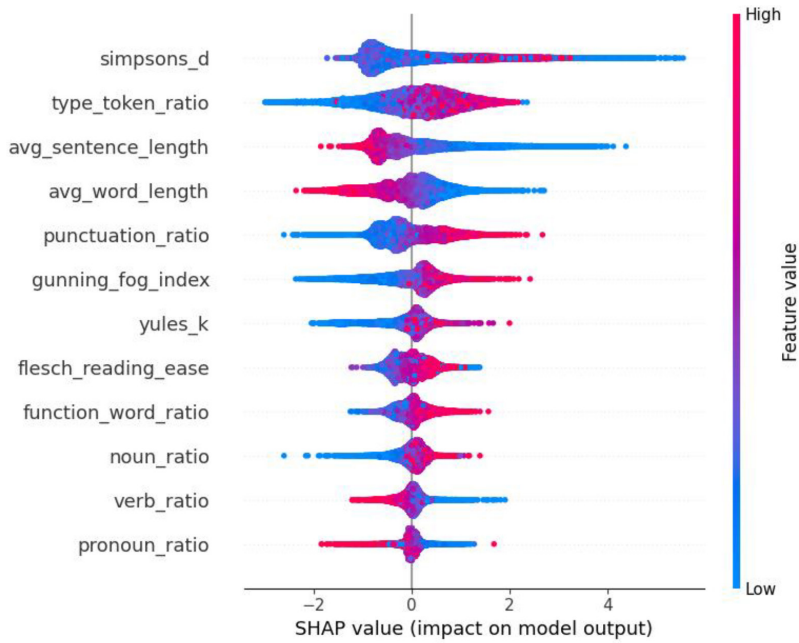

Fig. 8. SHAP Analysis for Gradient Boosting

Conclusions. This study demonstrates the efficacy of interpretable stylometric features in distinguishing between AI-generated and human-written text. Using a dataset of 30,000 balanced samples derived from ChatGPT, DeepSeek, and human responses, five machine learning models were evaluated on multiclass and binary classification tasks. Ensemble models, particularly Random Forest and Gradient Boosting, exhibited the highest performance, achieving macro F1-scores up to 0.84 and 0.86, respectively.

A key contribution of this work is the identification of stylometric indicators that remain relevant and effective even as generative AI continues to advance. Metrics such as Simpson's Diversity Index, type-token ratio, average sentence length, and punctuation frequency emerged as consistently informative across models. The SHAP analysis and feature importance rankings confirmed that interpretable linguistic features provide robust and meaningful signals for distinguishing text origin.

Furthermore, this research highlights the practical value of affordable dataset generation using open-source local models within the Ollama framework. This allows researchers to create diverse, labeled corpora without reliance on paid APIs or restricted datasets, supporting reproducibility and open experimentation.

Future work will extend this framework by evaluating transformer-based architectures such as RoBERTa, DistilBERT, and XLNet to compare deep contextual models with stylometric approaches. Additionally, domain-specific and multilingual datasets will be incorporated to assess generalization performance and robustness under varying linguistic conditions.

Ultimately, this study affirms that stylometry remains a powerful and interpretable tool for AI text detection. When combined with ensemble learning and lightweight feature engineering, it can deliver high-accuracy results with full transparency – making it well-suited for applications in education, publishing, and digital integrity assurance.

REFERENCES:

1. S. Gehrmann, H. Strobelt, and A. Rush, "GLTR: Statistical Detection and Visualization of Generated Text" in Proc. ACL: System Demonstrations, Florence, Italy, 2019, pp. 111–116. doi: 10.18653/v1/P19-3019

2. E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, and C. Finn, "DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature" arXiv preprint arXiv:2301.11305, 2023.

2. A. Akram, "An empirical study of AI generated text detection tools" arXiv:2310.01423, 2023.

3. C. Opara, "StyloAI: Distinguishing AI-Generated Content with Stylometric Analysis" arXiv:2405.10129, 2024.

4. L. Mindner, T. Schlippe, and K. Schaaff, "Classification of Human- and AI-Generated Texts: Investigating Features for ChatGPT" in Proc. 4th Int. Conf. Artificial Intelligence in Education Technology (AIET), Nov. 2023, pp. 152–170.

5. A. Uchendu, T. Le, K. Shu, and D. Lee, "Authorship Attribution for Neural Text Generation" in Proc. EMNLP, Nov. 2020, pp. 8384–8395. doi: 10.18653/v1/2020.emnlp-main.673.

6. G. Huang, Y. Zhang, Z. Li, Y. You, M. Wang, and Z. Yang, "Are AI-Generated Text Detectors Robust to Adversarial Perturbations?" in Proc. 62nd Annu. Meet. Assoc. Comput. Linguistics (ACL), Aug. 2024, pp. 6005–6024

7. A. M. Sarvazyan et al., "Overview of AuTexTification at IberLEF 2023: Detection and attribution of machine-generated text in multiple domains" Proces. Leng. Nat., vol. 71, pp. 275–288, 2023.

8. G. Mikros, A. Koursaris, D. Bilianos, and G. Markopoulos, "AI-writing detection using an ensemble of transformers and stylometric features" in CEUR Workshop Proc., vol. 3496, pp. 142–153, 2023.

9. J. Zhang, H. Sun, K. Duan, X. Li, M. Zhang, Y. Liu, and M. Sun, "How Would GPT Behave? Towards Detecting AI-Generated Text via Phrase-Level Self-Diversity," arXiv preprint arXiv:2301.07597, 2023.