Pinkey-bw at PAN 2026: Robust Generative AI Authorship Detection using RoBERTa and Stylometric Feature Fusion

Notebook for the PAN Lab at CLEF 2026

Bhavani Savalam[1] and Matti Wiegmann[2]

[1]Bauhaus-Universität Weimar, Germany
[2]University of Kassel, Germany

Abstract

The rapid advancement of large language models has significantly increased the difficulty of distinguishing between human-written and AI-generated text. Modern generative systems produce fluent and coherent content that often resembles authentic human writing styles, creating challenges for digital text forensics, academic integrity, misinformation detection, and authorship verification [6, 1]. In this work, we present a robust generative AI authorship detection framework developed for the PAN 2026 Voight-Kampff shared task at CLEF. Our approach combines contextual representations from RoBERTa with handcrafted stylometric features in a fusion architecture designed to capture both semantic and stylistic characteristics of text. To improve robustness against stylistic manipulation and adversarial paraphrasing, we introduce a stylometric obfuscation and data augmentation pipeline, motivated by the realistic authorship verification setting of the PAN Voight-Kampff task [2, 1].

Keywords

Generative AI Detection, Authorship Verification, Stylometry, RoBERTa, Transformer Models, Feature Fusion, Obfuscation, Data Augmentation

1. Introduction

The emergence of large language models such as GPT, LLaMA, Claude, Gemini, and Mistral has transformed natural language generation. These models can generate fluent text across academic writing, creative writing, conversational responses, and technical documentation. However, this progress also creates risks related to plagiarism, misinformation, impersonation, and synthetic content detection. Previous work has shown that humans often struggle to reliably distinguish AI-generated language from human-written language [6].

Detecting AI-generated text has become increasingly difficult because modern language models can imitate human writing styles with high accuracy. Earlier approaches often relied on statistical signals such as perplexity, entropy, token frequency, or compression-based features. Recent detection methods have explored probability curvature, log-rank information, and cross-perplexity signals for zero-shot machine-generated text detection [9, 8, 10, 7]. However, these approaches can be sensitive to changes in model families, domains, and text distributions.

The PAN 2026 Voight-Kampff Generative AI Detection shared task focuses on distinguishing human-written text from AI-generated text under realistic conditions. The task builds on earlier PAN and ELOQUENT editions of the Voight-Kampff authorship verification task [2, 1]. In this setting, generated texts may imitate specific human authors and may also be modified using paraphrasing or obfuscation techniques. Therefore, robust detection systems must learn semantic and stylistic patterns that generalize beyond shallow surface cues.

In this work, we propose a hybrid framework combining RoBERTa-based contextual embeddings with handcrafted stylometric features. Stylometric and authorship-verification methods have a long tradition in computational authorship analysis [11], and recent neural text generation studies further show that authorship-related signals can be useful for machine-generated text detection [5]. We also

CLEF 2026 Working Notes, 21–24 September 2026, Jena, Germany
bhavani.savalam@uni-weimar.de (B. Savalam); matti.wiegmann@uni-kassel.de (M. Wiegmann)

introduce a stylometric obfuscation pipeline for data augmentation, inspired by the need for robustness in LLM detection and authorship verification [3].

2. Task Description

The PAN Voight-Kampff shared task focuses on binary classification between human-written and AI-generated text. The task has evolved across PAN and ELOQUENT editions to address increasingly realistic AI detection challenges [2, 1]. The dataset contains samples from humans and multiple generative models across different genres and domains.

The training data used in this work contains labels where label 0 represents human-written text and label 1 represents AI-generated text. The training set contains 14,606 AI-generated samples and 9,101 human-written samples. The validation set contains 2,312 AI-generated samples and 1,277 human-written samples.

3. Methodology

The proposed system combines contextual semantic representations obtained from a pre-trained RoBERTa encoder with handcrafted stylometric features in a late-fusion architecture. This design follows the idea that AI-generated text detection benefits from combining different signal types, including semantic, statistical, and stylistic information [5, 3, 7]. In addition, a stylometric obfuscation pipeline is introduced to generate augmented training samples that expose the classifier to stylistic variations while preserving the underlying semantic content.

The complete pipeline consists of five stages:

1. Text preprocessing
2. Stylometric obfuscation and data augmentation
3. Stylometric feature extraction
4. RoBERTa encoding
5. Feature fusion and classification

3.1. Text Preprocessing

The preprocessing stage prepares the input text for both the RoBERTa encoder and the stylometric feature extraction module. Unlike the obfuscation stage described in Section 3.3, preprocessing is not intended to modify the writing style of a document. Instead, it removes formatting inconsistencies and document-specific artefacts while preserving stylistic characteristics that are informative for authorship detection.

The preprocessing pipeline performs the following operations:

- **Whitespace normalization.** Multiple consecutive whitespace characters are collapsed into a single space.
- **Repeated punctuation cleanup.** Runs of repeated punctuation symbols are normalized while preserving punctuation itself.
- **Basic text cleaning.** Formatting inconsistencies are corrected before tokenization.

Capitalization patterns, punctuation usage, digit usage, and lexical characteristics are intentionally preserved because they are explicitly used during stylometric feature extraction. Consequently, preprocessing standardizes the input representation without removing stylistic information that may help distinguish human-written from AI-generated text.

3.2. Stylometric Obfuscation and Data Augmentation

The preprocessing stage prepares the input text for feature extraction and transformer encoding, whereas the obfuscation stage generates additional training samples through controlled stylistic modifications. The two stages therefore serve different purposes.

The PAN Voight-Kampff shared task evaluates systems under realistic conditions where AI-generated texts may be paraphrased or intentionally rewritten to conceal their origin [2, 1]. A detector trained only on the original training corpus may therefore rely on surface-level stylistic cues that are weakened or absent during evaluation. To improve robustness, we generate additional stylistic variants of the original training documents while preserving their semantic content. This augmentation strategy exposes the classifier to a wider variety of writing styles and encourages it to learn semantic and stylometric representations that remain informative after stylistic modifications.

The following obfuscation techniques were implemented:

Masking URLs, email addresses and user identifiers: Regular-expression patterns were used to identify URLs, email addresses, and user identifiers in the text. Each matched expression was replaced by a reserved placeholder token (<URL>, <EMAIL>, or <USER>) while leaving the surrounding sentence unchanged. Unlike stylometric transformations, this operation is intended to remove document-specific textual entities that are not representative of an author's writing style. Replacing these expressions prevents the model from exploiting accidental correlations between specific identifiers and the training labels, encouraging it to learn stylistic and semantic properties that generalize across documents.

Number normalization: Numerical values were normalized and masked to reduce dependency on exact number patterns. This operation is intended to reduce the chance that the classifier relies on specific numerical values that may be dataset-specific rather than stylistic.

Contraction expansion: Common English contractions (e.g., "can't" → "cannot" and "I'm" → "I am") were expanded into their canonical forms. This transformation introduces lexical variation while preserving the semantic meaning of the sentence. As a result, the model is exposed to different lexical realizations of the same content and becomes less dependent on surface-level wording.

Filler-word removal: Frequently occurring discourse fillers such as "really", "actually", "basically", and "perhaps" were removed with a predefined probability. These words often contribute to an individual's writing style while carrying limited semantic information. Removing them generates stylistic variants of the same document and encourages the classifier to rely on more stable semantic and structural characteristics rather than author-specific filler usage.

Controlled synonym substitution: Lexical variation was introduced using a manually constructed synonym dictionary implemented in the augmentation script. Only a small set of high-frequency content words with unambiguous semantic alternatives was included to minimize the risk of changing the meaning of the original text. During augmentation, each candidate word was replaced with its predefined synonym according to a fixed probability, while all other words remained unchanged. This conservative strategy increases lexical diversity while preserving the semantic content of the document.

The manually defined synonym dictionary contains replacements such as:

- *important* → *significant*
- *show* → *demonstrate*
- *method* → *approach*
- *use* → *utilize*

- *help* → *assist*
- *begin* → *start*

Sentence splitting and merging: Adjacent sentences were probabilistically merged, whereas longer sentences were occasionally divided into shorter segments. This transformation alters sentence-length distributions and discourse structure while preserving the underlying information content. Consequently, the classifier is exposed to a wider range of writing styles during training.

Function-word balancing: Selected discourse markers and function words, including "however", "therefore", and "moreover", were probabilistically inserted when grammatically appropriate. Function words are known to contribute to writing style while having relatively little influence on semantic content. Varying their usage produces stylistic diversity without substantially changing the meaning of the document.

The individual transformations were selected because they modify different stylistic dimensions of a document, including lexical choice, sentence structure, and function-word usage, while preserving the original semantic content. This augmentation strategy reduces the likelihood that the classifier overfits to a limited set of surface-level stylistic cues and instead encourages learning representations that remain informative when stylistic characteristics are altered through paraphrasing and obfuscation.

3.3. Stylometric Feature Extraction:

In addition to contextual embeddings obtained from RoBERTa, the proposed framework extracts a fixed set of handcrafted stylometric features from each document. Stylometry has been widely used in authorship analysis and authorship verification [11], and previous work on neural text generation has shown that authorship-related signals can be useful for distinguishing human-written from generated text [5].

The feature extractor computes the following nine stylometric features:

- Number of words
- Number of characters
- Average word length
- Average sentence length
- Punctuation count
- Uppercase ratio
- Digit ratio
- Space ratio
- Lexical diversity (type–token ratio)

These features describe complementary aspects of writing style, including document length, lexical richness, sentence complexity, punctuation usage, capitalization patterns, and numerical usage. Since the extracted stylometric features have different numerical ranges, the handcrafted feature vector was standardized using the `StandardScaler` implementation from scikit-learn. The scaler transforms each feature to have zero mean and unit variance based on the training data. This prevents features with large numerical magnitudes from dominating the optimization process and ensures that all stylometric features contribute comparably during fusion with the RoBERTa representation.

3.4. RoBERTa-based Fusion Model

RoBERTa-base was used as the main transformer encoder. Input texts were tokenized with truncation and padding using a maximum length of 512 tokens. The contextual representation from RoBERTa was combined with stylometric features using a fusion layer.

The overall architecture of the proposed fusion model is illustrated in Figure 1. The model consists of two parallel branches. The first branch extracts contextual semantic representations using the RoBERTa encoder, while the second branch extracts handcrafted stylometric features from the same input text. The projected stylometric representation is concatenated with the RoBERTa [CLS] embedding before binary classification.

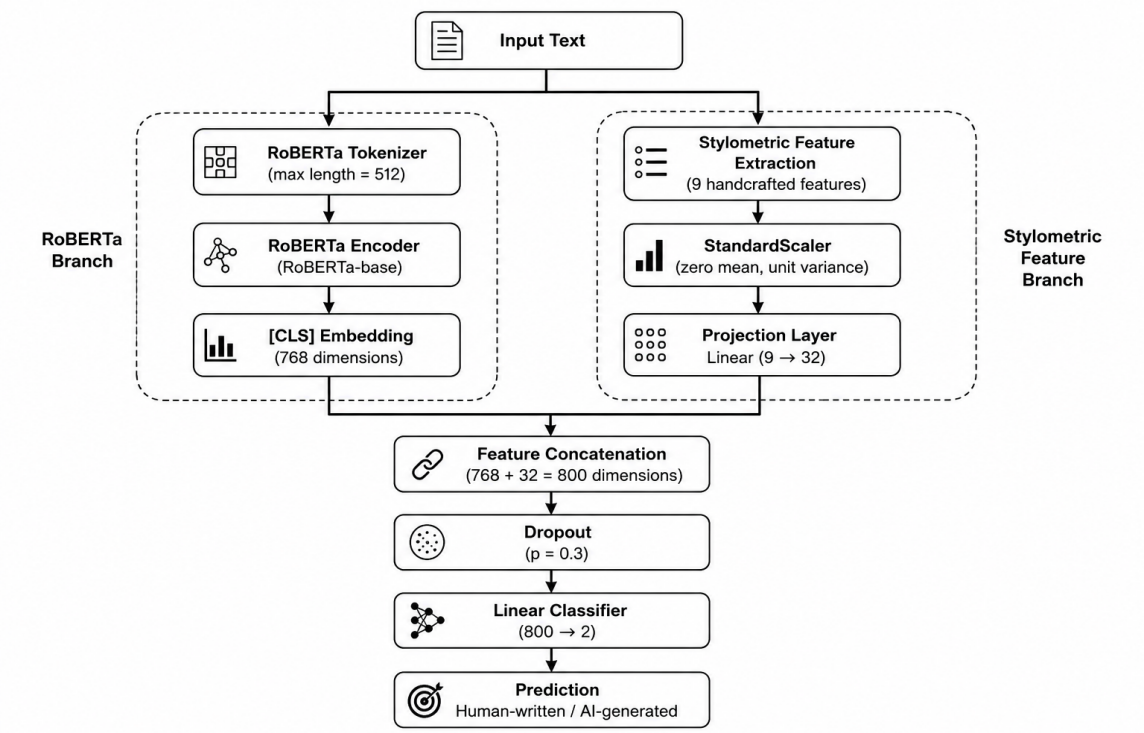

Figure 1: Architecture of the proposed RoBERTa–stylometric feature fusion model. The input text is processed through two parallel branches: the RoBERTa branch extracts contextual embeddings, while the stylometric branch extracts handcrafted linguistic features. The two representations are fused through feature concatenation and passed to a dropout layer followed by a linear classifier for binary prediction.

3.5. Training Strategy

The proposed model was trained as a supervised binary classifier using the AdamW optimizer with a learning rate of 2×10^{-5} and a weight decay of 0.01. Input documents were tokenized using the RoBERTa tokenizer with truncation and padding to a maximum sequence length of 512 tokens. Training was performed with a batch size of 8, and the model parameters were optimized using the cross-entropy loss function. During training, the stylometric feature vector was standardized using a `StandardScaler` fitted on the training set and subsequently fused with the contextual RoBERTa representation. The model producing the best validation performance was retained for inference.

4. Results and Discussion

4.1. Local Validation Experiments

To evaluate the effectiveness of the proposed stylometric obfuscation strategy, four experimental configurations were conducted using the datasets provided by the PAN 2026 shared task. The experiments investigate the influence of different training data variants while evaluating the model on both the original and obfuscated validation sets.

The first experiment used the original training dataset without augmentation. The second experiment trained the model using only stylometrically obfuscated training samples. In the third experiment, the original and obfuscated training datasets were combined to increase stylistic diversity during training. Finally, the best-performing mixed-data model was evaluated on an obfuscated version of the validation set to assess its robustness against stylistic modifications that preserve semantic content.

Table 1
Comparison of all four experimental settings across PAN metrics.

Setting	ROC-AUC	Brier	C@1	F1	F0.5u	Mean
Original Training Data	0.9990	0.9886	0.9886	0.9912	0.9875	0.9910
Obfuscated Training Data	0.9989	0.9845	0.9841	0.9878	0.9828	0.9876
Mixed Augmented Data	0.9994	0.9947	0.9947	0.9959	0.9951	0.9960
Obfuscated Validation Data	0.9994	0.9953	0.9953	0.9963	0.9965	0.9965

The results show that the proposed fusion architecture performs consistently well across all validation experiments, with ROC-AUC values above 0.998 in every configuration. Training exclusively on the obfuscated dataset resulted in a small reduction in performance compared with training on the original data, suggesting that the obfuscation process removes some stylistic cues that are useful for classification. However, the decrease remained limited, indicating that the contextual representations learned by RoBERTa remain highly informative.

The strongest performance on the original validation set was achieved when the model was trained on the mixed dataset containing both original and obfuscated samples. This suggests that combining clean and stylistically modified examples improves robustness without sacrificing classification accuracy. The mixed model also achieved strong performance on the obfuscated validation set, indicating that the augmentation strategy helps the model remain effective when the evaluation data contains stylistic modifications.

4.2. Official PAN Benchmark Evaluation

After validating the proposed approach locally, the final trained model was submitted to the official TIRA evaluation platform. Unlike the local validation experiments, these benchmark datasets contain hidden labels and were evaluated by the PAN organizers. Consequently, these results provide an unbiased assessment of the proposed method on unseen data generated under different conditions.

Table 2 reports the official benchmark results obtained on the PAN evaluation datasets.

Table 2
Official benchmark results obtained through the PAN 2026 TIRA evaluation platform.

Dataset	**ROC-AUC**	**Brier**	**C@1**	**F1**	**F0.5u**	**Mean**
PAN25 Test	0.941	0.928	0.928	0.946	0.971	0.943
PAN26 Test	0.660	0.504	0.504	0.511	0.716	0.579
Eloquent Test	0.700	0.438	0.438	0.572	0.770	0.584
PAN26 Smoke Test	0.969	0.971	0.971	0.968	0.987	0.973

The official benchmark evaluation shows that performance varies substantially across datasets. The strongest benchmark results were obtained on the PAN25 test set and the PAN26 smoke-test dataset. Performance was lower on the PAN26 hidden test set and the Eloquent benchmark, which suggests that these datasets contain different distributions or previously unseen generation conditions. The final system ranked 28th on the official PAN 2026 leaderboard, indicating competitive performance while also leaving room for improvement in cross-domain robustness.

5. Conclusion

This paper presented a RoBERTa–stylometric feature fusion framework for the PAN 2026 Voight-Kampff Generative AI Detection shared task. A stylometric obfuscation pipeline was introduced to generate augmented training data while preserving semantic content. Experimental results showed that combining original and obfuscated training samples produced the best overall validation performance and improved robustness to stylistic variations. The proposed fusion architecture effectively leveraged both contextual and handcrafted linguistic information for AI authorship detection. The final system was further evaluated on the official PAN benchmark datasets through the TIRA platform, where it ranked 28th on the PAN 2026 leaderboard. These results demonstrate the potential of combining transformer-based representations with stylometric information for robust AI-generated text detection. Future work will investigate stronger transformer models and more advanced augmentation strategies to improve generalization to unseen AI generators.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT and language tools to support grammar checking, paraphrasing, improving academic style, and organizing content. After using these tools and services, the authors reviewed and edited the content as needed and took full responsibility for the content of the publication.

References

[1] Janek Bevendorff, Matti Wiegmann, Jussi Karlgren, Luise Dürlich, Evangelia Gogoulou, Aarne Talman, Efstathios Stamatatos, Martin Potthast, and Benno Stein. Overview of the "Voight-Kampff" Generative AI Authorship Verification Task at PAN and ELOQUENT 2025. In Guglielmo Faggioli, Nicola Ferro, Paolo Rosso, and Damiano Spina (eds.), *Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum*, CEUR Workshop Proceedings, pages 3504–3534. CEUR-WS.org, 2025.

[2] Janek Bevendorff, Matti Wiegmann, Jussi Karlgren, Luise Dürlich, Evangelia Gogoulou, Aarne Talman, Efstathios Stamatatos, Martin Potthast, and Benno Stein. Overview of the "Voight-Kampff" Generative AI Authorship Verification Task at PAN and ELOQUENT 2024. In Guglielmo Faggioli, Nicola Ferro, Petra Galuščáková, and Alba García Seco de Herrera (eds.), *Working Notes of CLEF 2024 – Conference and Labs of the Evaluation Forum*, CEUR Workshop Proceedings, pages 2486–2506. CEUR-WS.org, 2024.

[3] Janek Bevendorff, Matti Wiegmann, Emmelie Richter, Martin Potthast, and Benno Stein. The Two Paradigms of LLM Detection: Authorship Attribution vs. Authorship Verification. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3762–3787. Association for Computational Linguistics, 2025.

[4] Janek Bevendorff, Xavier Bonet Casals, Berta Chulvi, Daryna Dementieva, Ashraf Elnagar, Dayne Freitag, Maik Fröbe, et al. Overview of PAN 2024: Multi-Author Writing Style Analysis, Multilingual Text Detoxification, Oppositional Thinking Analysis, and Generative AI Authorship Verification: Extended Abstract. In *Lecture Notes in Computer Science*, pages 3–10. Springer Nature Switzerland, 2024.

[5] Adaku Uchendu, Thai Le, Kai Shu, and Dongwon Lee. Authorship Attribution for Neural Text Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8384–8395. Association for Computational Linguistics, 2020. doi:10.18653/v1/2020.emnlp-main.674.

[6] Maurice Jakesch, Jeffrey T. Hancock, and Mor Naaman. Human Heuristics for AI-Generated

Language Are Flawed. *Proceedings of the National Academy of Sciences*, 120(11):e2208839120, 2023. doi:10.1073/pnas.2208839120.

[7] Abhimanyu Hans, Avi Schwarzschild, Valeriia Cherepanova, Hamid Kazemi, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Spotting LLMs with Binoculars: Zero-Shot Detection of Machine-Generated Text. arXiv:2401.12070, 2024.

[8] Jinyan Su, Terry Yue Zhuo, Di Wang, and Preslav Nakov. DetectLLM: Leveraging Log Rank Information for Zero-Shot Detection of Machine-Generated Text. arXiv:2303.12289, 2023.

[9] Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. DetectGPT: Zero-Shot Machine-Generated Text Detection Using Probability Curvature. arXiv:2301.11305, 2023.

[10] Guangsheng Bao, Yanbin Zhao, Zhiyang Teng, Linyi Yang, and Yue Zhang. Fast-DetectGPT: Efficient Zero-Shot Detection of Machine-Generated Text via Conditional Probability Curvature. arXiv:2310.05130, 2023.

[11] Moshe Koppel and Jonathan Schler. Authorship Verification as a One-Class Classification Problem. In *Proceedings of the Twenty-First International Conference on Machine Learning (ICML)*, pages 489–495, 2004.