Team SINAI at PAN 2026: Combining Stylometric Embeddings and IDF Features for Generative AI Text Detection

Notebook for the PAN Lab at CLEF 2026

César Espin-Riofrio[1,*], Arturo Montejo-Ráez[2], Janis Lara-Jama[1] and Jeremy Castro-Illescas[1]

[1]University of Guayaquil, Delta Av. s/n, Guayaquil, 090510, Ecuador
[2]University of Jaén, Las Lagunillas s/n, Jaén, 23071, Spain

Abstract

We describe the Sinai submission to Subtask 1 of the PAN 2026 Voight-Kampff Generative AI Detection shared task, a binary classification challenge covering three genres: fiction, news, and essays. As large language models produce increasingly fluent text, detecting AI-generated content requires signals that go beyond surface fluency. We approach this by combining mStyleDistance, a multilingual stylometric encoder that captures writing style independently of content, with binary IDF features that recover the lexical distributional cues stylometric representations deliberately suppress. The resulting classifier is trained end-to-end and further strengthened through probability calibration and an abstention mechanism designed to optimize the official ranking metric directly. On the official smoke-test evaluation, the system achieves a mean score of 0.968 with a ROC-AUC of 0.958.

Keywords

AI text detection, stylometric embeddings, IDF features, isotonic calibration

1. Introduction

The rapid proliferation of large language models has made it increasingly difficult to distinguish human-written text from AI-generated content. PAN 2026 frames this as a binary classification problem over three genres: fiction, news, and essays [1]. Each genre presents its own challenge: fiction rewards stylistic subtlety, news rewards lexical precision, and essays sit somewhere in between.

Most recent detection approaches rely either on neural classifiers built on top of large language models or on handcrafted feature pipelines, often struggling to generalize across genres or unseen generators. Rather than choosing between these extremes, we combine a multilingual stylometric encoder fine-tuned end-to-end with a compact IDF projection that captures lexical distributional cues complementary to the stylometric signal [2]. Post-training calibration and an abstention mechanism further refine the probability outputs to maximize the official ranking metric, yielding robust detection across the three genres of the shared task.

2. Related Work

Detecting AI-generated text has attracted growing attention as language models become more capable. Zero-shot methods such as Binoculars exploit perplexity contrasts to detect machine-generated text without labeled data, achieving strong results in single-model settings but degrading under model or domain shift [3].

CLEF 2026 Working Notes, September 21-24 2026, Jena, Germany
*Corresponding author.
cesar.espinr@ug.edu.ec (C. Espin-Riofrio); amontejo@ujaen.es (A. Montejo-Ráez); janis.laraj@ug.edu.ec (J. Lara-Jama); jeremy.castroi@ug.edu.ec (J. Castro-Illescas)
0000-0001-8864-756X (C. Espin-Riofrio); 0000-0002-8643-2714 (A. Montejo-Ráez); 0009-0000-1987-2042 (J. Lara-Jama); 0009-0007-8557-9259 (J. Castro-Illescas)

Supervised approaches address this limitation by fine-tuning transformer models on labeled corpora. TURINGBENCH [4] established a multi-model benchmark that revealed how detectors trained on one generator often fail on another, motivating more robust feature choices. The PAN shared tasks on generative AI authorship verification have provided a consistent evaluation framework across genres and models since 2024 [5, 6], with a TF-IDF linear SVM serving as a surprisingly competitive baseline (mean score 0.978 on the 2026 validation set). In this line, Espin-Riofrio et al. [2] explored classification-based approaches combining text features for AI-generated text detection, providing direct precedent for the hybrid architecture proposed here.

Stylometric methods offer a complementary perspective: rather than detecting statistical artifacts of generation, they capture writing style at a deeper level. Qiu et al. [7] introduced mStyleDistance, a multilingual encoder trained with contrastive learning to produce content-independent style representations that generalize across domains and languages. We build on this encoder by concatenating a sparse IDF projection that recovers the lexical signal that stylometric embeddings deliberately suppress. The resulting hybrid classifier is further extended with layer-wise learning rate decay, isotonic calibration, and test-time augmentation to improve both discrimination and probability calibration across genres.

3. System Description

3.1. Architecture

Our classifier, `StyleAIClassifierIDF`, processes each input through two parallel branches, as illustrated in Figure 1. The first encodes the text with mStyleDistance [7], an XLM-RoBERTa-based encoder trained with contrastive learning to produce content-independent style embeddings of 768 dimensions, fine-tuned end-to-end. The second transforms the same text with a static IDF vectorizer into a sparse 5000-dimensional binary vector, which then passes through a projection head (Linear(5000→256) → LayerNorm → GELU → Dropout) that reduces it to 256 dimensions, preventing the lexical vector from dominating the stylometric embedding during optimization. The concatenated representation [768 + 256] feeds a two-layer classifier: Linear(1024, 384) → GELU → Linear(384, 2).

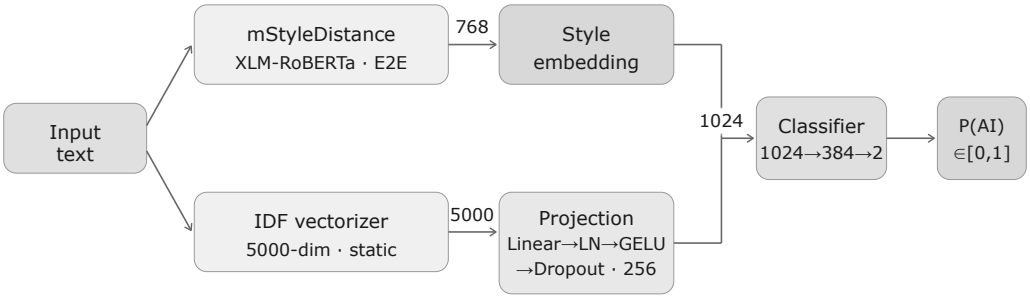

Figure 1: Architecture of `StyleAIClassifierIDF`. The stylometric branch (left) is fine-tuned end-to-end; the IDF branch (right) remains static. Both representations are concatenated before the final classifier.

3.2. IDF Configuration

The vectorizer is fitted once on the training set before any gradient updates and remains static throughout training and inference. We use `max_features=5000`, `ngram_range=(1,4)`, `binary=True`, and `min_df=2`, implemented via scikit-learn's `TfidfVectorizer` with `sublinear_tf=False`. With `binary=True`, term frequency is replaced by binary presence, so each feature reflects IDF-weighted term occurrence rather than raw frequency. This configuration outperformed standard TF-IDF particularly in the news genre.

3.3. Preprocessing

Each text undergoes minimal cleaning: URLs, user mentions, and hashtags are removed and repeated whitespace is normalized. The pipeline does not lowercase or strip punctuation, deliberately preserving the capitalization and punctuation patterns that carry stylistic information.

3.4. Training

We fine-tune with AdamW using layer-wise learning rate decay (LLRD, factor=0.85): the classification head receives the base learning rate, each transformer layer receives a rate decayed by 0.85 per layer from top to bottom, and the embedding layer receives the lowest rate. This prevents catastrophic forgetting of the pre-trained stylometric representations. We use cross-entropy loss with label smoothing 0.1, Mixup regularization (alpha=0.2, probability=0.3) in the embedding space, and a cosine schedule with 8% warmup. Training runs for up to 9 epochs with early stopping (patience 4) on validation F1-macro. Mixed precision is disabled due to incompatibility between mStyleDistance and BFloat16 on the G4 GPU; gradient checkpointing is enabled to fit training within available memory. Table 1 summarizes the main hyperparameters.

Table 1
Training hyperparameters.

Hyperparameter	Value
Learning rate (base)	6×10^{-6}
LLRD factor	0.85
Effective batch size	64 (16×4 accum.)
Max sequence length	256
Epochs (max)	9
Early stopping patience	4
Label smoothing	0.1
Mixup α	0.2
Warmup ratio	0.08
Gradient clip	1.0
Weight decay	0.01
Dropout	0.20
Hidden dim	384

3.5. Class Imbalance and Data Augmentation

The dataset presents genre-level class imbalance: essays and news show AI-to-human ratios around 4:1, while fiction is approximately balanced. We address this with a genre-stratified `WeightedRandomSampler` that equalizes the contribution of each genre-label combination within each mini-batch. Additionally, we augment only the human class in essays and news using three techniques: random word deletion, sentence shuffling, and text truncation. News receives a higher augmentation ratio (0.25 vs. 0.15 for essays), reflecting its higher false-positive rate observed during development (14.3% vs. 7.6% for essays).

3.6. Post-processing

Raw model probabilities are refined through three sequential steps. **First**, isotonic regression calibration is fitted via 5-fold out-of-fold cross-validation on the validation set to align confidence scores with empirical accuracy, avoiding data leakage between the calibrator and the held-out evaluation split. **Second**, test-time augmentation (TTA) generates three views of each input text via truncation at different ratios and averages their calibrated scores, reducing variance on borderline cases. **Third**, an abstention mechanism maps scores in the interval $[0.44, 0.56]$ to exactly 0.5, signaling uncertainty to

the evaluator; C@1 and $F_{0.5u}$ reward abstention over incorrect predictions, making this step directly beneficial for the mean score. The final decision threshold is fixed at 0.50 post-calibration.

Two systems were submitted to TIRA: *IDF only* (run `idf_concat`), which uses the architecture described above without any post-processing, and *IDF + post-processing* (run `silver-cabinet`), which extends it with the calibration, TTA, and abstention pipeline described in Section 3.6.

3.7. Experimental Setup

All experiments use the training and validation splits provided by the task organizers through TIRA [8]. The smoke-test partition (`pan26-generative-ai-detection-smoke-test-20260330-training`) serves as the primary benchmark reported here, since official test-set labels are not yet available at the time of writing. The dataset covers three genres: fiction, news, and essays, with texts labeled as human-written (0) or AI-generated (1) by a variety of large language models.

Models are trained on a single G4 GPU through Google Colab. Each test sample is processed in isolation, with no information leakage between samples. The isotonic calibrator and IDF vectorizer are fitted on the training and validation splits respectively, and remain static during inference.

The six official evaluation metrics are ROC-AUC, Brier score, C@1, F1-macro, $F_{0.5u}$, and their arithmetic mean, which is the primary ranking criterion. Submissions appear on the leaderboard under the group name `ce-ec`[1].

4. Results

Table 2 reports the official TIRA scores on the smoke-test partition for our two submitted systems. The IDF + post-processing system achieves a mean score of 0.968, with particularly strong $F_{0.5u}$ (0.987) and C@1 (0.971), reflecting the effectiveness of isotonic calibration and the abstention mechanism. The IDF only system, which shares the same architecture but omits the post-processing pipeline, reaches a mean score of 0.952, confirming that calibration and test-time augmentation contribute a consistent gain of 0.016 points.

Table 2
Official TIRA scores on the smoke-test partition.

System	ROC-AUC	Brier	C@1	F1	$F_{0.5u}$	Mean
IDF + post-processing (silver-cabinet)	0.958	0.959	0.971	0.968	0.987	**0.968**
IDF only (idf_concat)	0.990	0.955	0.941	0.938	0.938	0.952

The two systems present an interesting trade-off. IDF only achieves higher ROC-AUC (0.990 vs 0.958), indicating stronger raw discriminative power, while IDF + post-processing converts that signal into better binary decisions through calibration and abstention, yielding higher C@1, F1, and $F_{0.5u}$. Since the mean score weights all five metrics equally, the post-processing gains outweigh the ROC-AUC reduction.

Table 3 breaks down the contribution of each post-processing step on the validation set. Isotonic calibration is the most impactful component, with its removal causing the largest drop in mean score ($0.976 \longrightarrow 0.968$). Abstention and TTA contribute marginally in isolation but reinforce each other in the full pipeline.

Table 4 confirms the central claim of this work: removing the IDF projection causes a mean score drop of 0.120 points ($0.976 \longrightarrow 0.856$), demonstrating that lexical and stylometric features are strongly complementary rather than redundant.

For reference, the organizers report a TF-IDF SVM baseline with a mean score of 0.978 on the official validation set [9]; a direct comparison with our systems on the same partition will be possible once test-set results are released.

[1] A name change request to `SINAI` has been submitted to the task organizers.

Table 3
Post-processing ablation on the validation set (2,871 samples).

Configuration	ROC-AUC	Brier	C@1	F1	$F_{0.5u}$	Mean
Full system (calib + TTA + abstention)	0.979	0.974	0.972	0.977	0.975	0.976
Without abstention	0.979	0.974	0.972	0.977	0.975	0.975
Without TTA	0.972	0.976	0.975	0.981	0.974	0.976
Without isotonic calibration	0.977	0.967	0.964	0.973	0.959	0.968
Without post-processing	0.974	0.966	0.963	0.972	0.958	0.966

Table 4
IDF ablation on the validation set (2,871 samples). Post-processing applied to both systems.

System	ROC-AUC	Brier	C@1	F1	$F_{0.5u}$	Mean
IDF + post-processing (full)	0.979	0.974	0.972	0.977	0.975	0.976
mStyleDistance only (no IDF)	0.879	0.868	0.859	0.832	0.841	0.856

5. Error Analysis

Out of 2,871 validation examples, the system produces 36 false positives (1.3%) and 49 false negatives (1.7%), with no abstentions on this partition. Fiction concentrates most false negatives (40 of 49, 2.7% of the genre), while news shows the highest false positive rate (20 of 36, 2.3%). Essays are the most reliably classified genre, with only 5 false positives and 2 false negatives.

Among false negatives, Gemini 1.5 Pro and Llama 3.1 8B Instruct are the hardest generators to detect, accounting for 14 (11.6%) and 13 (10.6%) of their respective validation examples. Gemini-generated fiction approximates human stylistic variation closely enough to evade both the stylometric encoder and the IDF features. Llama 3.1 8B errors are more surprising given its smaller scale, suggesting this model produces outputs with atypically low lexical and stylistic divergence from human text. False positive scores are consistently high (mean=0.964), indicating the model is overconfident on these cases, while false negative scores cluster near zero (mean=0.238), suggesting the model assigns very low AI probability to these texts rather than being uncertain about them.

6. Conclusion

We presented the Sinai submission to the PAN 2026 Voight-Kampff Generative AI Detection task, combining a multilingual stylometric encoder with a binary IDF projection for generative AI text detection across three genres. Ablation experiments confirm the central contribution: removing the IDF projection causes a mean score drop of 0.120 points (0.976 $\longrightarrow$ 0.856), demonstrating that lexical and stylometric features are strongly complementary rather than redundant. Beyond the architecture itself, the post-processing pipeline of isotonic calibration, test-time augmentation, and abstention proved essential to translate discriminative power into competitive ranking scores, with isotonic calibration identified as the most impactful component.

Among generators, Gemini 1.5 Pro and Llama 3.1 8B Instruct remain the hardest to detect, with false negative rates of 11.6% and 10.6% respectively, both concentrated in the fiction genre. Future work will explore genre-specific calibration and ensemble approaches as potential paths toward closing this gap.

Acknowledgments

This work is funded by the Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia - Funded by EU – NextGenerationEU within the framework of the project Desarrollo Modelos ALIA. This work has also been partially supported by

Project CONSENSO (PID2021-122263OB-C21) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR, Project ROMANET (CERV-2024-CHAR-LITI-101215052), funded by the European Union under the Citizens, Equality, Rights and Values programme, Project HEART-NLP-UJA (PID2024-156263OB-C21) and project VERITAS-H (AIA2025-163322-C64) funded by MICIU/AEI/10.13039/501100011033 and by ERDF/EU, Project GALENO-IA (DGP_PIDI_2024_00852) funded by the European Union under the Smart Specialization Strategy Program for Sustainability in Andalusia (S4Andalucía) 2021–2027. This research was also supported by Project FCI-036-2023 at the University of Guayaquil, Ecuador.

Declaration on Generative AI

During the preparation of this work, the author used Claude (Anthropic) in order to assist in drafting and editing this paper. The author reviewed and edited all content as needed and takes full responsibility for the publication's content.

References

[1] J. Bevendorff, M. Fröbe, A. Greiner-Petter, A. Jakoby, M. Mayerl, P. Nakov, H. Plutz, M. Potthast, B. Stein, M. N. Ta, Y. Wang, E. Zangerle, in: M. Hagen, M. Potthast, B. Stein, P. Schaer, E. Zangerle, A. Barrón-Cedeño, A. G. S. de Herrera, E. S. Salido, S. MacAvaney, J. M. Struß (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2026.

[2] C. Espin-Riofrio, et al., Detection of automatically generated texts: A classification-based approach and text features, in: Communications in Computer and Information Science, Springer, 2026.

[3] A. Hans, A. Schwarzschild, V. Cherepanova, H. Kazemi, A. Saha, M. Goldblum, J. Geiping, T. Goldstein, Spotting LLMs with binoculars: Zero-shot detection of machine-generated text, in: Proceedings of the 41st International Conference on Machine Learning (ICML 2024), PMLR, 2024.

[4] A. Uchendu, Z. Ma, T. Le, R. Zhang, D. Lee, TURINGBENCH: A benchmark environment for turing test in the age of neural text generation, in: Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, 2021, pp. 2001–2016.

[5] J. Bevendorff, et al., Overview of the "Voight-Kampff" generative AI authorship verification task at PAN and ELOQUENT 2025, in: Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, 2025, pp. 3504–3534.

[6] J. Bevendorff, R. R. Gunti, J. Karlgren, M. Fröbe, M. Potthast, B. Stein, Overview of the third "Voight-Kampff" Generative AI / LLM Detection Task at PAN and ELOQUENT 2026, in: A. Barrón-Cedeño, A. G. S. de Herrera, E. S. Salido, S. MacAvaney, J. M. Struß (Eds.), Working Notes of CLEF 2026 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, 2026.

[7] J. Qiu, J. Zhu, A. Patel, M. Apidianaki, C. Callison-Burch, mStyleDistance: Multilingual style embeddings and their evaluation, in: Findings of the Association for Computational Linguistics: ACL 2025, Association for Computational Linguistics, Vienna, Austria, 2025.

[8] M. Fröbe, M. Wiegmann, N. Kolyada, B. Grahm, T. Elstner, F. Loebe, M. Hagen, B. Stein, M. Potthast, Continuous Integration for Reproducible Shared Tasks with TIRA.io, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2023, pp. 236–241. URL: https://link.springer.com/chapter/10.1007/978-3-031-28241-6_20. doi:10.1007/978-3-031-28241-6_20.

[9] J. Bevendorff, D. Dementieva, M. Fröbe, B. Gipp, A. Greiner-Petter, J. Karlgren, M. Mayerl, P. Nakov, A. Panchenko, M. Potthast, A. Shelmanov, E. Stamatatos, B. Stein, Y. Wang, M. Wiegmann, E. Zangerle, Overview of PAN 2025: Generative AI Authorship Verification, Multi-Author Writing Style Analysis, Multilingual Text Detoxification, and Generative Plagiarism Detection, in: Exper-

imental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2025), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2025.