Team aimoment at PAN 2026: Content-Agnostic Contrastive Stylometry for Multi-Author Writing Style Analysis

Notebook for the PAN Lab at CLEF 2026

Momen Ibrahim[1,*], Nagwa El-Makky[1] and Marwan Torki[1]

[1]Computer and Systems Engineering Department, Faculty of Engineering, Alexandria University, Alexandria, Egypt

Abstract

Detecting author boundaries in multi-author documents requires models that distinguish *writing style* independently from *topical content* — a separation that becomes critical when consecutive sentences share the same subject matter. We present **ContraStyle**, a system built around this principle: every design decision is aimed at suppressing topical shortcuts and amplifying pure stylistic signals. The core component is a DeBERTa-v3 pairwise classifier trained with a composite loss that combines cross-entropy, Supervised Contrastive Learning, R-Drop, and a novel **Content-Agnostic Contrastive Objective (CACO)** — an InfoNCE loss that uses style-change pairs from the *same document* as hard negatives, explicitly preventing the model from exploiting shared vocabulary. This is complemented by a LightGBM classifier over handcrafted stylometric features and an SBERT-based sequential profiler (SSPC) that models the trajectory of style across the full document. All components are trained per difficulty on data spanning PAN 2022–2026. On the official PAN 2026 test set, our best submission achieves F1-macro of **1.000 / 0.741 / 0.773** for easy / medium / hard.

Keywords

multi-author writing style analysis, style change detection, DeBERTa, stylometric features, contrastive learning, R-Drop, ensemble

1. Introduction

The Multi-Author Writing Style Analysis task at PAN 2026 [1, 2] asks participants to detect *style-change boundaries* in documents authored by multiple writers. Formally, given a document $D = (s_1, s_2, \ldots, s_n)$, the system must output a binary label $y_i \in \{0, 1\}$ for each consecutive pair (s_i, s_{i+1}), where $y_i = 1$ indicates a change in writing style between the two sentences.

The task is decomposed into three sub-tasks of increasing difficulty: **Easy**—documents span multiple topics, providing a topic-shift signal; **Medium**—documents share a common topic with only $\approx 4.4\%$ positive pairs (extreme class imbalance); **Hard**—strictly same-topic documents where pure stylometric discrimination is required.

The central challenge escalates from easy to hard: the easier the topic signal is to exploit, the less the model needs to understand style. A system that works well on easy by tracking vocabulary shifts will fail on hard precisely *because* it learned the wrong signal. The unifying design principle of ContraStyle is therefore **content-agnostic style learning**: every component is designed to suppress topical shortcuts and amplify the stylistic signal that persists when topic is held constant. This manifests at three levels—contrastive training objectives that use same-document hard negatives, handcrafted features that measure writing behaviour rather than semantic content, and a sequential model that captures the trajectory of style across the full document. All signals are fused through a calibrated ensemble with per-difficulty decision thresholds.

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany

*Corresponding author.

es-momen.ibrahim2019@alexu.edu.eg (M. Ibrahim); nagwamakky@alexu.edu.eg (N. El-Makky); mtorki@alexu.edu.eg (M. Torki)

https://github.com/Momen-Ibrahim-Fawzy/ (M. Ibrahim)

0009-0001-7287-0705 (M. Ibrahim); 0000-0002-9571-7543 (N. El-Makky); 0000-0002-6149-1718 (M. Torki)

Table 1
Training and validation pair counts (after deduplication and windowed pair generation).

Difficulty	Train pairs	Val pairs	Pos. rate
Easy	630,984	132,951	30.5%
Medium	1,453,481	303,326	4.4%
Hard	783,249	146,162	21.0%

2. Related Work

Multi-author writing style detection has been a recurring challenge at PAN since 2023 [3]. Prior work has explored siamese networks, BERT-based pairwise classifiers, and graph-based models. Top-performing systems in PAN 2024 employed DeBERTa fine-tuned on windowed sentence pairs with Virtual Softmax for sharper decision boundaries [4, 5]. We build on this approach by adopting DeBERTa with Virtual Softmax as our core pairwise classifier and extending it with contrastive and consistency objectives specifically designed to suppress topical shortcuts.

Pre-trained language models such as DeBERTa-v3 [6] have demonstrated strong performance on document-level tasks due to their disentangled attention mechanism. We adopt DeBERTa-v3-base as our backbone encoder, leveraging its disentangled attention for fine-grained style discrimination at the sentence-pair level. Supervised Contrastive Learning (SCL) [7] improves representation learning by pulling same-class embeddings together, which is beneficial for style discrimination. We incorporate SCL into our composite loss specifically to shape the projection space for content-agnostic style clustering, ensuring that same-author pairs are embedded close together regardless of topic.

R-Drop [8] reduces over-fitting by applying a bidirectional KL divergence between two dropout-augmented forward passes of the same input, acting as a consistency regulariser. We use R-Drop as a consistency regulariser targeting the instability that topic-driven representations exhibit under dropout perturbation — a complementary mechanism to CACO's explicit hard-negative construction. Sentence-BERT [9] enables efficient computation of semantic similarity features. SBERT embeddings serve dual roles in our system: as semantic distance features in the LightGBM stylometric component (Section 4.2) and as the input representation for the SSPC sequential profiler (Section 4.3). Sentence transformers have also been applied to the related task of authorship verification, demonstrating their effectiveness for capturing writing style beyond semantic similarity [10].

3. Data

We combine training data from all available PAN editions (2022–2026) [1, 2, 5] for each difficulty level after SHA-256 deduplication at the document level. Although the underlying texts have changed substantially across editions — particularly in 2026, where topic control is stricter — multi-year pooling serves three purposes. First, it increases the diversity of writing styles and boundary patterns, which is critical for contrastive losses (especially CACO) that rely on in-batch diversity for meaningful hard negatives. Second, exposure to older editions with less controlled topic conditions forces the model to be robust to varying degrees of topical overlap, rather than overfitting to a single year's topic-control regime. Third, per-difficulty training and threshold calibration on the 2026 validation set mitigate any distribution shift introduced by older data. The final training corpus spans approximately 24K–30K documents per difficulty. Table 1 summarises the resulting pair counts after windowed pair generation.

Windowed pair construction. Rather than encoding each boundary as a single sentence pair (s_i, s_{i+1}), we concatenate a *context window* of W sentences on each side:

$$\text{left} = s_{i-W} \oplus \cdots \oplus s_i, \quad \text{right} = s_{i+1} \oplus \cdots \oplus s_{i+W+1} \tag{1}$$

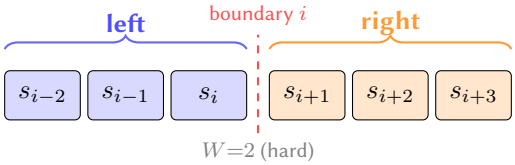

Figure 1: Windowed pair construction for boundary i with $W{=}2$ (hard). Three sentences to the left (blue) form **left**; three to the right (orange) form **right**. With $W{=}1$ (easy/medium), two sentences appear on each side.

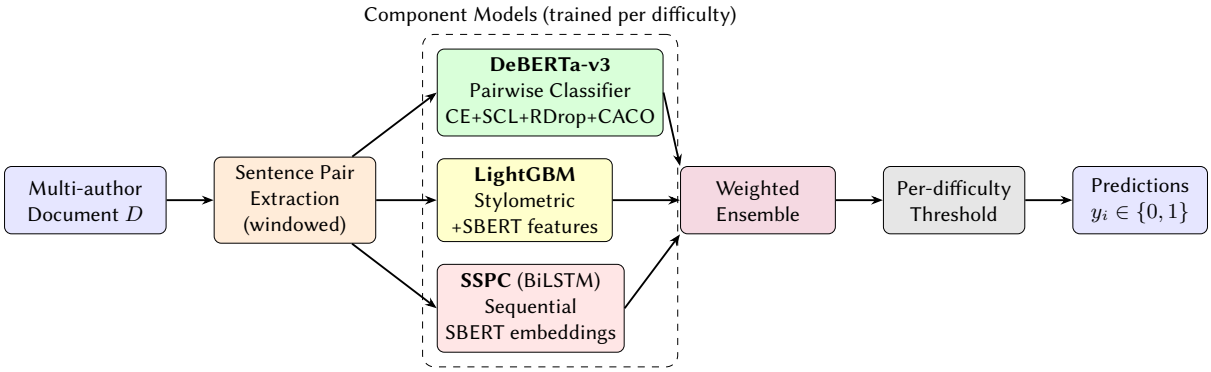

Figure 2: Overview of the ContraStyle architecture. Windowed sentence pairs are fed to three independently trained component models; their probability scores are blended and thresholded per difficulty level.

where $\oplus$ denotes newline-separated concatenation. The window is tokenized as a sentence pair:

$$\texttt{[CLS]}\ \text{left}\ \texttt{[SEP]}\ \text{right}\ \texttt{[SEP]}$$

We use $W{=}1$ for easy and medium (two sentences per side, 256 token budget) and $W{=}2$ for hard (three sentences per side, 384 token budget). The larger window for hard provides more stylometric context when topic content is held constant. Figure 1 illustrates this construction.

4. System Description

Figure 2 provides an overview of the system. Three component models — each designed to capture a different facet of content-agnostic style — independently score each sentence-pair boundary. The DeBERTa classifier operates at the *local pair level*, using contrastive losses with same-document hard negatives to prevent topic leakage. The LightGBM classifier operates over *surface stylometric signals* (word-length distributions, function-word rates, punctuation patterns) that are inherently topic-independent. The SSPC BiLSTM operates at the *document sequence level*, modelling how the style trajectory evolves across the entire document. Their probability scores are linearly blended and thresholded per difficulty level via grid search on the validation set.

4.1. DeBERTa-v3 Pairwise Classifier

We fine-tune DeBERTa-v3-base [6] for all difficulty levels as a pairwise style-change classifier. The model is presented with a windowed sentence pair and produces a classification over the `[CLS]` token representation via a linear head with N_{out} outputs.

Model architecture. The DeBERTa encoder produces a contextualised `[CLS]` representation $\mathbf{h} \in \mathbb{R}^d$. Two parallel heads consume $\mathbf{h}$:

- **Classifier head**: $\mathbf{W}_c\mathbf{h} \in \mathbb{R}^{N_{\text{out}}}$, used for the cross-entropy loss and final prediction.

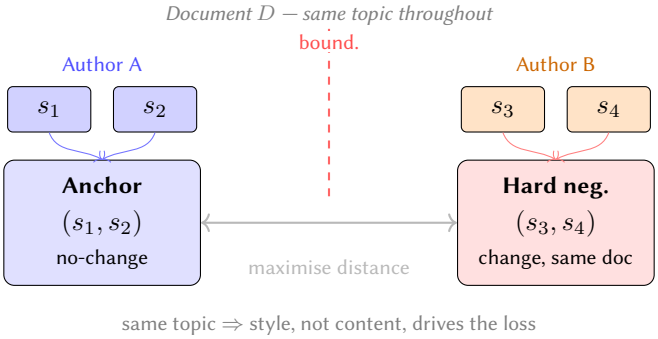

Figure 3: CACO hard-negative construction. The anchor (no-change pair, Author A) and the hard negative (style-change pair, Author B) share the same topical context because they come from the same document. InfoNCE maximises their embedding distance, preventing the encoder from exploiting shared vocabulary.

- **Projection head**: a two-layer MLP projecting $\mathbf{h}$ to a 128-dimensional L_2-normalised embedding $\mathbf{z}$ for the contrastive losses.

Virtual Softmax. We employ the Virtual Softmax technique [4], where the classifier head outputs three logits (N_{out}=3) even though there are only two true classes. The third (phantom) class acts as a regulariser that forces the model to push the real class logits higher for confident predictions. At inference, only the first two logits are used and re-normalised via softmax.

Composite loss function. Training minimises a composite loss:

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \alpha_{\text{SCL}}\,\mathcal{L}_{\text{SCL}} + \alpha_{\text{RD}}\,\mathcal{L}_{\text{RD}} + \alpha_{\text{CACO}}\,\mathcal{L}_{\text{CACO}} \tag{2}$$

with α_{SCL}=0.3, α_{RD}=0.3, α_{CACO}=0.2. Each component targets a distinct failure mode of the content-agnostic objective. Cross-entropy provides the direct supervised classification signal but does not shape the embedding geometry. SCL [7] fills this gap by pulling same-class embeddings together in the projection space, producing compact and well-separated style clusters that generalise better to unseen author combinations. R-Drop [8] penalises predictions that vary across dropout masks, targeting the instability of topic-driven representations that change with surface perturbations rather than with genuine style differences. Finally, CACO directly addresses the core challenge of topic–style confounding: without it, the model can minimise cross-entropy by exploiting shared vocabulary within same-topic documents rather than learning genuine stylistic differences.

Cross-entropy loss ($\mathcal{L}_{\text{CE}}$). Weighted cross-entropy with class weights proportional to the inverse positive rate per difficulty and label smoothing ε=0.05.

Supervised Contrastive Loss ($\mathcal{L}_{\text{SCL}}$) [7]. Applied on projection embeddings $\mathbf{z}$, pulling same-class pairs together and pushing different-class pairs apart with temperature τ=0.1.

R-Drop ($\mathcal{L}_{\text{RD}}$) [8]. Each batch is forwarded twice with different dropout masks, yielding logit pairs $(\mathbf{l}_1, \mathbf{l}_2)$. The symmetric KL divergence

$$\mathcal{L}_{\text{RD}} = \tfrac{1}{2}\left[D_{\text{KL}}(p_1 \| p_2) + D_{\text{KL}}(p_2 \| p_1)\right] \tag{3}$$

acts as a consistency regulariser, reducing the variance of predictions under different dropout masks.

Content-Agnostic Contrastive Loss ($\mathcal{L}_{\text{CACO}}$). For medium and hard difficulties, same-topic documents make it easy for the model to conflate topic and style signals. We address this with an InfoNCE loss [11] that uses *in-batch hard negatives*: for each no-change anchor pair, the hard negatives are style-change pairs from the *same document* (same topic, different author). This forces the encoder to rely on stylistic rather than topical cues. Figure 3 illustrates the construction of hard negatives.

Table 2
Per-difficulty DeBERTa training configuration.

Setting	Easy	Medium	Hard
Model	base	base	base
Window W	1	1	2
Max tokens	256	256	384
Batch size	32	16	16
Grad. accum.	4	8	8
Effective batch	128	128	128
Learning rate	2e-5	1.5e-5	1e-5
LLRD decay	0.9	0.9	0.9
FGM ε	0.5	0.5	0.5
EMA decay	0.9995	0.9995	0.9995
Focal loss γ	–	2.0	2.0
Max epochs	15	25	30
Positive rate	30.5%	4.4%	21.0%

Training configuration. All models use AdamW [12] with Layer-wise Learning Rate Decay (LLRD, decay factor 0.9), a linear warm-up for 6% of training steps, and cosine decay. An effective batch size of 128 is maintained via gradient accumulation (Table 2).

We apply three additional regularisation techniques. **FGM** [13] (Fast Gradient Method, ε=0.5) perturbs word embeddings in the gradient direction before the optimiser step, improving generalisation via adversarial training. **EMA** (Exponential Moving Average, decay 0.9995) maintains a shadow copy of weights used for validation and saving the best model. **Focal Loss** [14] (γ=2.0) is applied for medium and hard difficulties to down-weight easy negatives under severe class imbalance.

4.2. Stylometric Feature Classifier (LightGBM)

We extract 108 handcrafted stylometric features for each sentence pair and train a LightGBM [15] gradient-boosted decision tree classifier. Features fall into four categories:

- **Lexical statistics**: average word length, type-token ratio, function-word rate, punctuation density, sentence length (characters/words).
- **Syntactic features**: POS-tag unigram distributions (noun, verb, adjective, adverb, pronoun ratios), dependency depth estimates.
- **Character-level patterns**: uppercase ratio, digit ratio, special-character distributions.
- **Semantic distance**: cosine similarity between SBERT (all-mpnet-base-v2) sentence embeddings [9] for each sentence in the pair, plus the L_2 distance and the element-wise absolute difference.

Feature differences between the left and right sides of a boundary are computed as the L_1 distance vector. LightGBM is trained with up to 1 000 estimators with early stopping on validation F1-macro. Figure 4 shows the top-12 features by gain for the hard difficulty; easy and medium exhibit a similar ranking.

4.3. Sequential Style-Profiling Classifier (SSPC)

To capture long-range style dynamics within a document, we train a BiLSTM classifier (SSPC) that processes the sequence of frozen SBERT embeddings for all sentences in a document. The BiLSTM learns to detect style transitions from the trajectory of sentence representations in the embedding space.

For each document, SBERT embeddings $\mathbf{e}_1, \ldots, \mathbf{e}_n$ are fed to the BiLSTM. The concatenated forward and backward hidden states at each boundary position are passed to a linear classifier producing

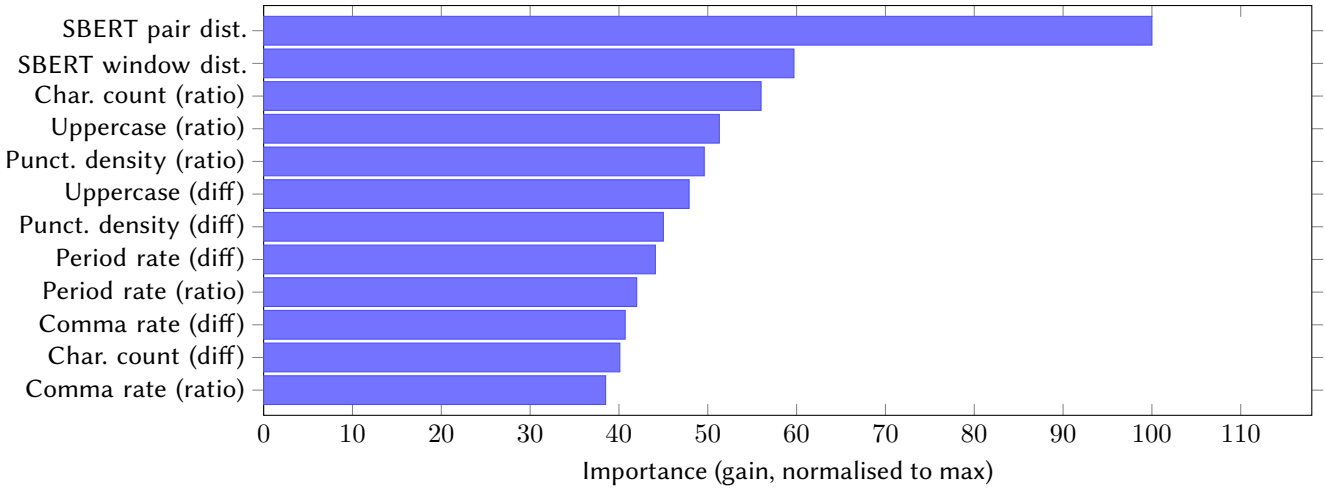

Figure 4: Top-12 LightGBM feature importances (gain, normalised to the maximum) for hard difficulty. SBERT-based distances are most discriminative; punctuation rates and character counts follow. All 108 features are topic-agnostic.

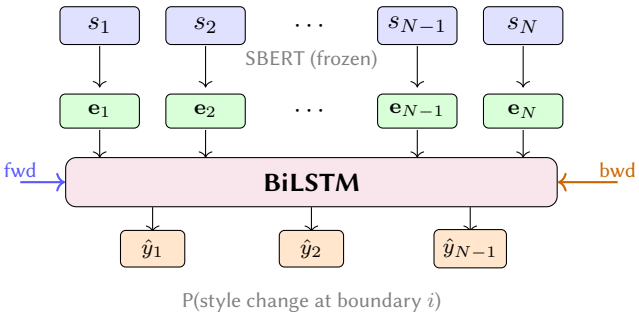

Figure 5: SSPC architecture. Frozen SBERT sentence embeddings $\mathbf{e}_1, \ldots, \mathbf{e}_N$ are fed to a BiLSTM. At each boundary i, the concatenated forward–backward hidden state is projected to produce the style-change probability $\hat{y}_i$, capturing long-range style dynamics across the full document.

style-change probabilities. The model is trained with binary cross-entropy on document-level batches (all pairs within one document per step), using a cosine annealing learning-rate schedule.

Per-difficulty SSPC hyperparameters differ in hidden size (256 for easy/medium; 512 for hard) and dropout (0.3 vs. 0.5) to reflect the increasing difficulty of the style-change signal. Figure 5 illustrates the architecture.

4.4. Ensemble and Threshold Calibration

The three component probability scores are combined via a weighted linear blend:

$$p_{\text{final}} = w_1 \cdot p_{\text{DeBERTa}} + w_2 \cdot p_{\text{LGBM}} + w_3 \cdot p_{\text{SSPC}} \tag{4}$$

Blend weights (w_1, w_2, w_3) and a per-difficulty decision threshold τ are jointly optimised by grid search on the validation set to maximise F1-macro. DeBERTa receives the highest weight ($\approx$0.70) given its consistently superior individual performance.

The threshold τ is critical for medium difficulty due to its severe class imbalance (4.4% positive rate), where the optimal threshold is substantially lower than 0.5.

Table 3
Validation F1-macro per component model and difficulty level.

Component	Easy	Medium	Hard
DeBERTa-v3	0.9972	0.8156	0.7868
LightGBM	0.9874	0.6604	0.5192
SSPC	0.9862	0.7408	0.6448

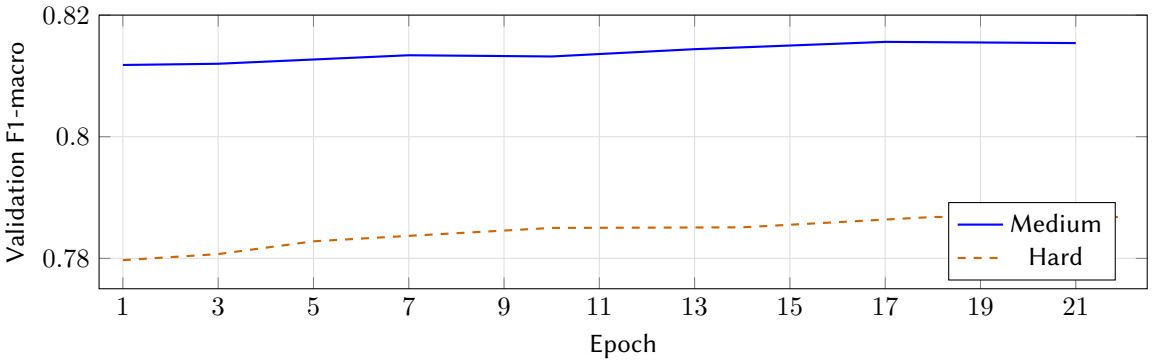

Figure 6: DeBERTa-v3 validation F1-macro per epoch for medium (blue, 21 epochs) and hard (orange dashed, 22 epochs). Easy reaches >0.997 in two epochs and is omitted. Both curves improve steadily; the hard model benefits from the larger training budget ($\leq$30 epochs).

5. Experimental Setup

Training and Software. Training was performed in PyTorch with HuggingFace Transformers and the `sentence-transformers` library. Docker containers were used for reproducible submission via the TIRA platform [16].

Evaluation metric. The task is evaluated by F1-macro averaged over the two classes (change / no-change) for each difficulty separately.

6. Results

Validation set performance. Table 3 reports the best validation F1-macro for each component model and difficulty level. DeBERTa dominates across all settings, with SSPC providing complementary sequential context especially for medium and hard. LightGBM contributes strongly on easy where stylometric signals are distinct but lags on medium due to severe class imbalance. Figure 6 shows the DeBERTa validation F1-macro by epoch for medium and hard; both difficulties show steady, monotone improvement with gradual convergence.

Component ablation. To quantify each component's contribution beyond the individual scores in Table 3, we evaluate four ensemble configurations on the validation set with per-configuration threshold re-calibration via grid search (Table 4). DeBERTa alone already achieves near-ceiling performance on easy (0.997) and strong baselines on medium (0.816) and hard (0.787), confirming that the composite contrastive loss is the primary engine of content-agnostic style discrimination. Adding the stylometric LightGBM classifier provides small but consistent gains (+0.0016 on medium, +0.0018 on hard), indicating that surface-level writing behaviour — punctuation rates, word-length distributions — captures style information complementary to DeBERTa's contextualised representations. The largest single-component improvement comes from SSPC: its document-level style trajectory raises medium from 0.816 to 0.823 (+0.0077), the biggest gain in the table. This is consistent with the design rationale:

Table 4
Component ablation on the validation set (F1-macro). Decision thresholds are re-calibrated per configuration via grid search.

Configuration	Easy	Medium	Hard
DeBERTa	0.9972	0.8156	0.7868
DeBERTa + LightGBM	0.9974	0.8172	0.7886
DeBERTa + SSPC	0.9974	0.8233	0.7882
Full ensemble	0.9979	0.8199	0.7886

Table 5
Official test-set F1-macro scores for our two submissions on the PAN 2026 held-out test set.

Submission	Easy	Medium	Hard
Run 1 (best)	**1.000**	**0.741**	**0.773**
Run 2	0.999	0.741	0.765

Table 6
PAN 2026 leaderboard: best run per team (F1-macro on the official test set).

Rank	Team	Easy	Medium	Hard
1	stylefusion	**1.000**	**0.767**	**0.775**
2	**aimoment (ours)**	**1.000**	**0.741**	**0.773**
3	ulille	0.995	0.712	0.769
4	spcrc	0.993	0.722	0.757
5	holovchyk-ind	0.972	0.699	0.733
6	datateam-tug	0.918	**0.768**	**0.780**
7	stitze	0.999	0.536	0.441
8	rheaveaura	0.997	0.558	0.580

SSPC models sequential style dynamics across the full document, capturing global author-transition patterns that neither the local pairwise classifier nor the surface feature classifier can observe. On medium, the DeBERTa+SSPC pair (0.823) slightly outperforms the full three-component ensemble (0.820), suggesting that adding LightGBM marginally dilutes the sequential signal under severe class imbalance (4.4% positive rate). On easy and hard, the full ensemble achieves the best or tied-best F1-macro (0.998 and 0.789), confirming that all three views contribute when the class distribution is more balanced.

Official test set results. Table 5 reports the F1-macro scores on the official PAN 2026 held-out test set for our two submissions. Both runs use the same three trained component models (DeBERTa-v3, LightGBM, SSPC) but differ in ensemble calibration. Run 1 applies a three-component blend with weights and decision threshold jointly optimised on the validation set via grid search. Run 2 uses DeBERTa and LightGBM only (SSPC weight set to zero), with independently selected per-difficulty thresholds. Including SSPC in the ensemble (Run 1) yields a gain on hard (0.773 vs. 0.765) while medium remains unchanged (0.741) and easy is effectively identical (1.000 vs. 0.999). Table 6 compares the best run per team across all participants.

7. Discussion

Difficulty-specific observations. On the official test set, our system achieves perfect F1-macro (1.000) for easy, confirming that topic-shift signals are strong enough that even stylometric features suffice when topics differ. Medium remains the most challenging difficulty: despite achieving validation

F1 of 0.816, our test-set score of 0.741 reveals a generalisation gap, likely due to the extreme class imbalance (4.4% positive pairs) and topic-domain shift between the training and test corpora. Per-difficulty threshold calibration (τ=0.12 for medium) is essential here. On hard, the larger window context (W=2), CACO loss, and Focal Loss together provide effective content-agnostic style discrimination.

The content-agnostic principle in practice. Each regularisation technique serves the same overarching goal: prevent topical signals from masking stylistic ones. R-Drop adds a consistency constraint between two dropout-augmented views of the same input, penalising predictions that depend on surface patterns that vary under perturbation — precisely the kind of instability that topic-driven representations exhibit. CACO goes further by constructing hard negatives from style-change pairs within the *same document*: the model is directly penalised for assigning similar embeddings to sentences that differ in author but share the same topical context. Focal Loss (γ=2.0, applied to medium and hard) reduces the influence of the trivially-negative majority class, forcing the model to attend to the stylistically informative minority. Together, these objectives do not merely regularise the model — they *define* what the model should attend to.

Limitations of the content-agnostic assumption. CACO constructs hard negatives from style-change pairs within the same document, relying on the assumption that intra-document pairs share the same topical context. This assumption holds most strongly for hard-difficulty documents, which are strictly same-topic, but is weaker for medium, where topic shifts can occur within a single document. Similarly, older PAN editions (2022–2024) had less controlled topic conditions, so some CACO negatives may pair different-topic sentences, diluting the content-agnostic signal. CACO therefore provides a best-effort suppression of topical shortcuts via intra-document structure, but does not guarantee complete topic exclusion. The remaining gap between validation and test performance on medium (0.816 vs. 0.741) may partly reflect residual topic leakage that CACO cannot fully eliminate.

Multi-view fusion and multi-year data. Combining three complementary views (local pairwise, surface stylometric, sequential trajectory) provides robustness: when one view is misled by topic content, the others anchor the ensemble. Combining PAN 2022–2026 editions substantially increased training set size (>1.4M pairs for medium), which is critical for contrastive losses that rely on in-batch diversity for meaningful hard negatives.

8. Conclusion

We presented ContraStyle, our system for the PAN 2026 Multi-Author Writing Style Analysis task. Our three-component ensemble combines a fine-tuned DeBERTa-v3 pairwise classifier (composite loss: CE + SCL + R-Drop + CACO), a LightGBM stylometric classifier, and a BiLSTM sequential profiler over SBERT embeddings. Adversarial training (FGM), Exponential Moving Average weights, Layer-wise LR Decay, and Focal Loss further strengthen the DeBERTa component.

On the official PAN 2026 test set our best submission achieves F1-macro of 1.000 / 0.741 / 0.773 for easy / medium / hard.

Our primary insight is that the hard difficulty requires content-agnostic training signals: R-Drop (reducing dropout variance), CACO (in-batch same-document hard negatives), and Focal Loss together push the model to focus on writing style rather than shared topic content. The main remaining challenge is medium difficulty, where the 4.4% positive rate causes a significant validation-to-test generalisation gap. Future work will explore DeBERTa-v3-large for medium and hard difficulties and improved calibration techniques for highly imbalanced settings.

Codes are publicly available at: https://github.com/Momen-Ibrahim-Fawzy/ContraStyle-PAN2026

Acknowledgments

We thank the PAN 2026 organizers for providing the evaluation infrastructure via TIRA [16] and the benchmark datasets.

Declaration on Generative AI

During the preparation of this work, the author used Claude (Anthropic) in order to: assist with code development and writing assistance. After using this tool, the author reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] E. Zangerle, M. Mayerl, M. Potthast, B. Stein, Overview of the Multi-Author Writing Style Analysis Task at PAN 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.

[2] 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.

[3] 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: Experimental 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.

[4] F. N. Chen, et al., Team baker at PAN 2024: DeBERTa with virtual softmax for multi-author writing style analysis, in: Working Notes of CLEF 2024, 2024. [Update with full author list and page numbers once proceedings are available].

[5] E. Zangerle, M. Mayerl, M. Potthast, B. Stein, Overview of the multi-author writing style analysis task at PAN 2024, in: Working Notes of CLEF 2024, 2024. [Update with full bibliographic details once proceedings are available].

[6] P. He, J. Gao, W. Chen, DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing, in: The Eleventh International Conference on Learning Representations, 2023. URL: https://openreview.net/forum?id=sE7-XhLxHA.

[7] P. Khosla, P. Tian, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, D. Krishnan, Supervised contrastive learning, in: Advances in Neural Information Processing Systems, volume 33, 2020, pp. 18661–18673. URL: https://proceedings.neurips.cc/paper/2020/hash/d89a66c7c80a29b1bdbab0f2a1a94af8-Abstract.html.

[8] L. Wu, J. Hu, L. Lyu, J. Li, W. He, F. Wu, R-drop: Regularized dropout for neural networks, in: Advances in Neural Information Processing Systems, volume 34, 2021, pp. 10567–10581. URL: https://proceedings.neurips.cc/paper/2021/hash/5a66b9200f29ac3fa0ae244cc2a51b39-Abstract.html.

[9] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using siamese BERT-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2019, pp. 3982–3992. URL: https://aclanthology.org/D19-1410/.

[10] M. Ibrahim, A. Akram, M. Radwan, R. Ayman, M. Abd-El-Hameed, M. Torki, Enhancing authorship verification using sentence-transformers, in: Working Notes of CLEF 2023, 2023. URL: https://ceur-ws.org/Vol-3497/paper-216.pdf.

[11] A. van den Oord, Y. Li, O. Vinyals, Representation learning with contrastive predictive coding, in: Advances in Neural Information Processing Systems (NeurIPS), 2018. ArXiv:1807.03748.

[12] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, in: International Conference on Learning Representations, 2019. URL: https://openreview.net/forum?id=Bkg6RiCqY7.

[13] T. Miyato, A. M. Dai, I. Goodfellow, Adversarial training methods for semi-supervised text classification, in: International Conference on Learning Representations, 2017. URL: https://openreview.net/forum?id=r1X3g2_xl.

[14] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2980–2988. URL: https://openaccess.thecvf.com/content_iccv_2017/html/Lin_Focal_Loss_for_ICCV_2017_paper.html.

[15] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, T.-Y. Liu, LightGBM: A highly efficient gradient boosting decision tree, in: Advances in Neural Information Processing Systems, volume 30, 2017. URL: https://proceedings.neurips.cc/paper/2017/hash/6449f44a102fde848669bdd9eb6b76fa-Abstract.html.

[16] 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. doi:10.1007/978-3-031-28241-6_20.