Team StyleFusion at PAN 2026: DeBERTa Meets Stylometry

Notebook for the PAN Lab at CLEF 2026

Adam Skurla[1,2]

[1]Faculty of Information Technology, Brno University of Technology, Brno, Czechia
[2]Kempelen Institute of Intelligent Technologies, Bratislava, Slovakia

Abstract

This notebook presents StyleFusion, our system for the PAN 2026 Multi-Author Writing Style Analysis task. The task focuses on detecting sentence-level style-change boundaries. StyleFusion combines a DeBERTa-v3-large classifier with a LightGBM stylometric model, and the final predictions are fused with logistic regression. Separate models and thresholds are used for the Easy, Medium, and Hard settings. On the official test set, the system achieved F1 scores of 1.000, 0.767, and 0.775 for the Easy, Medium, and Hard settings, respectively, corresponding to the highest mean F1 score among the submitted systems shown in the official leaderboard. The analysis shows that surface-level stylometric features are most useful in the Easy setting, while similarity and contextual features become more important in the more difficult settings.

Keywords

style change detection, stylometry, DeBERTa, model fusion, PAN

1. Introduction

Style change detection (referred to as *SCD* from now on) aims to identify positions in a document where the writing style changes. This problem is important for the analysis of multi-author documents, forensic linguistics, authorship analysis, and the study of text revisions [1]. It has also become more relevant with the increasing use of automatic writing tools, because documents can be produced through multiple authors, later edits, or mixed human-machine writing processes.

The PAN 2026 Multi-Author Writing Style Analysis shared task, organized as part of the PAN lab at CLEF 2026, focuses on sentence-level style change detection [2, 3]. The input is a document split into sentences, and the goal is to decide for each pair of neighboring sentences whether a change of author or writing style occurred. The task is divided into three difficulty settings: easy, medium, and hard [3]. In easier settings, style changes may also correlate with topic changes. In the hard setting, topic differences are limited, so systems need to rely more on stylometric differences between neighboring sentences [2].

In this notebook, we describe StyleFusion, our system for the PAN 2026 Multi-Author Writing Style Analysis task. StyleFusion combines two complementary branches. The first branch uses DeBERTa-v3-large [4] as a transformer-based sentence-pair classifier with local context around each sentence boundary. The second branch uses explicit stylometric features that capture lexical, syntactic, punctuation, and discourse differences between neighboring sentences, and classifies them with LightGBM [5]. The output probabilities of both branches are combined with logistic regression. We train separate models and tune separate decision thresholds for each difficulty setting.

The contributions of the proposed approach are as follows:

- we propose a **hybrid SCD system** that combines a transformer-based sentence-pair classifier with explicit stylometric features;
- we model each sentence boundary with **local left and right context**, rather than classifying isolated sentence pairs only;
- we train **separate models for the easy, medium, and hard settings** to account for different levels of topical variation;

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
adam.skurla@kinit.sk (A. Skurla)
0009-0000-9337-2804 (A. Skurla)

- we use **probability-level fusion with logistic regression** to combine contextual and stylometric evidence.

The source code of StyleFusion is publicly available at https://github.com/AdamSkurla/pan26-stylefusion.

2. Background

Stylometry studies measurable properties of writing style and is commonly used in authorship attribution, plagiarism detection, and forensic linguistics [6, 7]. Recent work includes both traditional feature-based approaches and neural architectures that model writing style through contextual representations [8]. These approaches capture different aspects of writing style and are often used together.

SCD extends authorship analysis to transitions within a document. Given a sequence of sentences, the task is to predict whether a style change occurs between neighboring sentences, without prior knowledge about the possible authors. The task has been organized as part of the PAN shared competition since 2016 [9, 10, 11, 2] and is summarized in recent surveys [1, 12, 13].

This trend is also visible in recent SCD systems. Earlier approaches combined transformer embeddings with traditional classifiers and feature-based methods [14]. More recent systems fine-tune transformer models directly for the task, including RoBERTa [15, 16, 17], ELECTRA [18, 16, 17], DeBERTa [18, 16], SqueezeBERT [16], ALBERT [15], and Llama [19]. These models capture strong contextual information and often achieve strong performance, but their predictions are usually difficult to interpret.

For this reason, some recent work combines neural representations with explicit stylometric features. Kumarage and Liu [8] combined PLM representations with stylometric features for authorship attribution, while Boriceanu and Băltoiu [20] combined transformer embeddings with handcrafted structural and stylometric features for PAN 2025 style change detection.

Inspired by these hybrid approaches, we combine transformer-based contextual modeling with explicit stylometric features for sentence-level style change detection.

3. System Overview

StyleFusion processes each document as a sequence of n sentences and independently classifies each of the $n - 1$ neighboring sentence boundaries as either a style change or not. Figure 1 shows the overall system pipeline. The system contains two parallel branches that produce a probability score for each sentence boundary.

The first branch is a *transformer branch* based on the *DeBERTa-v3-large* model, which has been used in recent SCD systems. Its goal is to capture contextual and semantic information around each sentence boundary. The second branch is a *stylometric branch* based on a LightGBM classifier. This branch models stylistic differences between neighboring sentences using explicit stylometric features extracted from the text.

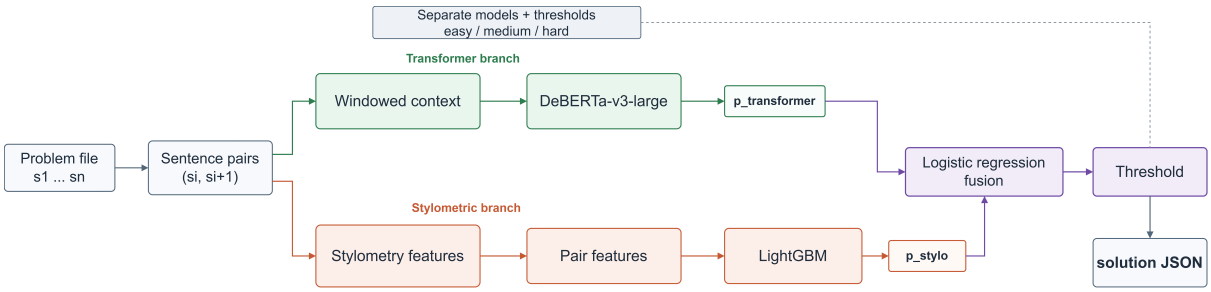

Figure 1: Overview of the StyleFusion pipeline.

The outputs of both branches are combined using a logistic regression fusion model, which produces the final prediction score. Separate decision thresholds are then selected for each difficulty setting (easy, medium, and hard). As a result, separate model configurations are trained for each task setting.

3.1. Preprocessing

The input consists of a text file with one sentence per non-empty line. A document with n sentences produces $n - 1$ boundary examples, where each example represents the transition between sentences s_i and s_{i+1}. Training labels are read from the `changes` field in the ground-truth JSON file. During inference, the predicted binary sequence is written to the corresponding `solution-problem-X.json` file.

For the transformer branch, each boundary is represented as a sentence pair with additional local context. The left side combines s_{i-1} with s_i, while the right side combines s_{i+1} with s_{i+2}. Boundaries at the beginning or end of the document use only available neighboring sentences. This local context helps the model capture stylistic continuity around each sentence boundary without processing the entire document.

3.2. Transformer Branch

The transformer branch fine-tunes the `microsoft/deberta-v3-large` model [4] for sentence-boundary classification. Each input consists of a left and right sentence context around the target boundary. The tokenized sentence pair is processed by the DeBERTa encoder, and the hidden states of all non-padding tokens are aggregated using mean pooling into a single 1024-dimensional representation. This representation is passed through a lightweight classification head with dropout regularization, dimensionality reduction, ReLU activation, and layer normalization, producing a scalar prediction score that is converted into the transformer probability score $p_{\text{transformer}}$ using a sigmoid function.

Training is optimized for the imbalanced binary classification setting, where non-change boundaries are more frequent than style-change boundaries. We therefore combine Focal Loss ($\gamma = 2.0$) with class-weighted binary cross-entropy. The model is trained with AdamW using a learning rate of 1×10^{-5}, weight decay of 0.1, and linear warmup during the first 10% of training steps. The maximum input length is limited to 256 tokens. Training runs for up to 20 epochs with early stopping based on validation macro-F1, and separate checkpoints are selected for the easy, medium, and hard settings.

3.3. Stylometric Branch

The stylometric branch extracts both sentence-level and pairwise features. Sentence-level features capture lexical, structural, syntactic, punctuation, and discourse-related properties of individual sentences. Pairwise features describe stylistic differences between both sides of the boundary. Table 1 summarizes the main feature categories used in the LightGBM branch.

For the medium and hard settings, the LightGBM feature vector also includes two contextual similarity features: cosine similarity and L2 distance between mean-pooled DeBERTa embeddings of neighboring sentences. These features are mainly useful in more difficult settings, where surface-level stylometric differences are often limited. Neighboring sentences may have similar length, punctuation, or POS structure while still differing in local context and formulation. Embedding similarity therefore complements handcrafted stylometric features with a finer contextual signal.

The final fixed-length feature vector is used as input to a LightGBM classifier [5], trained separately for each difficulty setting. Hyperparameters are optimized with Optuna using validation macro-F1 as the objective metric. The classifier outputs the stylometric probability score p_{stylo}.

3.4. Fusion

The fusion stage combines the two scalar probability scores, $[p_{\text{transformer}}, p_{\text{stylo}}]$, using a logistic regression classifier with balanced class weights. To avoid data leakage, the fusion model was trained on a

Table 1
Overview of handcrafted stylometric feature categories used in the LightGBM branch. The Easy setting uses 95 handcrafted features, while the Medium and Hard settings additionally include two contextual similarity features, resulting in 97 features in total.

Category	#	Representative features
Lexical	6	type-token ratio, rare-word ratio, contraction rate, character entropy
Pronouns	4	first-, second-, and third-person pronoun rates
Punctuation	7	comma, semicolon, question mark, ellipsis, dash rates
POS & Syntax	10	POS ratios, dependency depth, passive voice, modality, tense
Discourse & Style	4	discourse markers, hedge words, intensifiers, sentiment
Informal Markers	4	URLs, emojis, capitalization, informal markers
Sentence Structure	9	dependency distance, formality score, TTR, nominalization rate
Pairwise Features	53	feature differences, sentence length ratio, character and POS n-gram similarity, Jaccard overlap, function-word cosine, punctuation profile, contextual similarity

separate calibration split comprising 15% of the official PAN training data. This subset was not used for training either the DeBERTa or stylometric branches and served exclusively for fusion calibration.

After training, the decision threshold τ is selected on the validation set by searching values in the range $\tau \in [0.10, 0.89]$ with a step size of 0.01. The threshold that maximizes validation macro-F1 is selected independently for each difficulty setting.

The final binary prediction for each boundary is computed as

$$\hat{y} = \mathbf{1}[f(p_{\text{transformer}}, p_{\text{stylo}}) \geq \tau].$$

The resulting binary sequence is then serialized into the PAN output format as `{"changes": [...]}`.

4. Results and Analysis

We evaluate the proposed system from two perspectives. First, we report validation performance of the final fused model across all difficulty settings. Second, we analyze the stylometric LightGBM branch to identify which handcrafted stylistic features have the strongest influence on model decisions.

4.1. Validation Results

In this section, we analyze the validation results of the final StyleFusion system. We focus on the differences between the individual difficulty settings and on the effect of decision-threshold tuning.

Table 2 shows the results after model fusion with logistic regression. We use macro-F1 as the main evaluation metric because the number of style-change boundaries differs across the difficulty settings. We also report precision and recall for the style-change class.

The Easy setting achieves almost perfect results across all metrics, which suggests that the system can reliably detect strong stylistic transitions between neighboring text segments. Performance decreases

Table 2
Validation results of the final StyleFusion system. Precision and recall are reported for the style-change class.

Setting	Pairs	Accuracy	Macro-F1	P(change)	R(change)
Easy	117435	1.00	1.00	1.00	1.00
Medium	278988	0.96	0.76	0.50	0.57
Hard	127380	0.84	0.77	0.62	0.68

more clearly in the Medium setting, where the largest drop appears in precision for the style-change class. This indicates a higher number of false positive predictions and suggests that the boundaries between styles are less distinct than in the Easy setting. The Hard setting remains difficult mainly because the differences between neighboring segments are often subtle. Despite the lower accuracy, the system maintains a relatively balanced precision and recall and achieves a macro-F1 score similar to the Medium setting.

Because the fused probabilities were not equally calibrated across all settings, we tuned a separate decision threshold for each difficulty level. Table 3 compares the default threshold $\tau = 0.50$ with the threshold selected on the validation data.

Table 3
Effect of decision-threshold tuning on validation results. Precision and recall are reported for the style-change class.

Setting	τ	Macro-F1	P(change)	R(change)	Pred. changes
Easy	0.50	1.00	1.00	1.00	35542
Easy	0.21	1.00	1.00	1.00	35550
Medium	0.50	0.69	0.30	0.75	30167
Medium	0.89	0.76	0.50	0.57	13749
Hard	0.50	0.77	0.62	0.68	29532

Threshold tuning has only a minimal effect in the Easy setting, which suggests that the predicted probabilities are already well separated with the default threshold. The largest effect appears in the Medium setting. Increasing the threshold from 0.50 to 0.89 substantially reduces the number of predicted style changes, which improves precision and also increases the macro-F1 score. In the Hard setting, the default threshold already provides a relatively balanced trade-off between precision and recall, so additional tuning does not lead to a noticeable improvement.

4.2. Feature Analysis

The LightGBM branch allows us to analyze which explicit stylistic features are used by the system during prediction. For model interpretation, we use gain importance and SHAP values. Gain importance measures how much a feature contributes to reducing the training objective through decision splits, while SHAP values estimate the contribution of individual features to specific predictions. Table 4 presents the most important features for each difficulty setting. The direction column is based on differences in average feature values between style-change and no-change boundaries in the validation data. The reported gain values therefore represent the relative importance of features within a LightGBM model and not a direct contribution to the final F1 score.

The importance of individual features differs across the difficulty settings. In the Easy setting, the model relies mainly on differences in sentence length, word length, and syllable usage. These features are effective when stylistic differences between neighboring segments are more visible. In the Medium setting, similarity-based features and tense-related features become more important. Lower character 4-gram similarity and smaller word overlap are more often associated with style changes.

In the Hard setting, contextual similarity features play the largest role. DeBERTa cosine similarity has

Table 4
Top LightGBM features by gain importance per difficulty setting.

Setting	Feature	Gain (%)	Direction
Easy	sentence length ratio	30.7	lower for change
Easy	difference in average syllables per word	17.3	higher for change
Easy	character 4-gram cosine similarity	13.0	lower for change
Easy	difference in word count	6.9	higher for change
Easy	difference in average word length	4.3	higher for change
Medium	character 4-gram cosine similarity	8.1	lower for change
Medium	difference in past-tense rate	7.1	higher for change
Medium	difference in present-tense rate	6.4	higher for change
Medium	DeBERTa cosine similarity	5.0	lower for change
Medium	Jaccard word overlap	4.5	lower for change
Hard	DeBERTa cosine similarity	14.9	lower for change
Hard	character 4-gram cosine similarity	10.0	lower for change
Hard	difference in past-tense rate	8.4	higher for change
Hard	difference in present-tense rate	7.2	higher for change
Hard	difference in contraction rate	5.2	higher for change

the highest gain importance, and lower similarity is more frequently linked to style changes. Differences in tense usage, character 4-gram similarity, and contraction rate also remain important. Similar patterns can be observed in the SHAP analysis. In the Easy setting, predictions are influenced mainly by sentence and word length features, while in the Medium and Hard settings the feature contributions are more balanced. Overall, the results suggest that different difficulty settings rely on different types of stylistic signals, which supports the use of separate stylometric models.

5. Official Results

Table 5 presents the official PAN 2026 test results obtained from the TIRA experimentation platform [21]. Our team, *StyleFusion*, achieved the best overall performance with a mean F1 score of 0.847 and ranked first among all submissions. The Easy setting was nearly saturated, with several teams reaching almost perfect performance, which indicates that this setting was not strongly discriminative. In contrast, the main performance differences appeared in the Medium and Hard settings, where our system achieved competitive and balanced results (0.767 and 0.775 F1, respectively). Although the *datateam-tug* team obtained slightly higher scores on the Medium and Hard settings individually, its lower Easy score reduced the overall mean performance

Table 5
Official PAN 2026 test results. For each team, the best run by mean F1 over the three task settings is shown.

Rank	Team	Easy	Medium	Hard	Mean
1	**StyleFusion (ours)**	**1.000**	0.767	0.775	**0.847**
2	aimoment	1.000	0.741	0.773	0.838
3	uille	0.995	0.712	0.769	0.825
4	sprc	0.993	0.722	0.757	0.824
5	datateam-tug	0.918	**0.768**	**0.780**	0.822
6	holovchyk-ind	0.972	0.699	0.733	0.801
7	rheaveaura	0.997	0.558	0.580	0.712
8	stitze	0.999	0.536	0.441	0.659
9	maik-test-3-30	0.410	0.489	0.441	0.447

6. Conclusion

In this paper, we presented StyleFusion, a hybrid system for sentence-level style change detection. The system combines a DeBERTa-v3-large transformer branch with a stylometric LightGBM branch and merges their outputs using logistic regression. Separate models and decision thresholds are used for the Easy, Medium, and Hard settings.

The results show that the Easy setting is almost completely solved, while the Medium and Hard settings remain more challenging. The feature analysis also shows that the importance of stylometric signals changes across the difficulty levels. In the Easy setting, the model relies mainly on surface-level differences such as sentence and word length. In the Medium and Hard settings, similarity-based and contextual features become more important. Future work could explore more advanced fusion strategies, improved probability calibration, a broader set of stylometric features, and a dedicated ablation study of the proposed left/right contextual sentence-boundary modeling.

Acknowledgments

This work was partially supported by the European Union under the Horizon Europe project AI-CODE, GA No. 101135437, and partially by *LorAI – Low Resource Artificial Intelligence*, a project funded by Horizon Europe under GA No.101136646. This work was also supported by the Ministry of Education, Youth and Sports of the Czech Republic through the e-INFRA CZ (ID:90254).

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] A. S. Alsheddi, M. E. B. Menai, Writing style change detection: state of the art, challenges, and research opportunities, Artificial Intelligence Review 58 (2025) 401.

[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] 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.

[4] P. He, X. Liu, J. Gao, W. Chen, Deberta: Decoding-enhanced bert with disentangled attention, arXiv preprint arXiv:2006.03654 (2020).

[5] 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, Advances in neural information processing systems 30 (2017).

[6] H. Sayoud, Deep learning vs. conventional approaches in stylometry: An authorship attribution study, in: Proceedings of the 2024 7th International Conference on Information Science and Systems, 2024, pp. 158–162.

[7] J. Savoy, Machine learning methods for stylometry, Cham: Springer (2020).

[8] T. Kumarage, H. Liu, Neural authorship attribution: Stylometric analysis on large language models, in: 2023 International conference on cyber-enabled distributed computing and knowledge discovery (cyberc), IEEE, 2023, pp. 51–54.

[9] J. Bevendorff, I. Borrego-Obrador, M. Chinea-Ríos, M. Franco-Salvador, M. Fröbe, A. Heini, K. Kredens, M. Mayerl, P. Pęzik, M. Potthast, et al., Overview of pan 2023: Authorship verification, multi-author writing style analysis, profiling cryptocurrency influencers, and trigger detection: Condensed lab overview, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2023, pp. 459–481.

[10] E. Zangerle, M. Mayerl, M. Potthast, B. Stein, Overview of the multi-author writing style analysis task at pan 2024., in: CLEF (Working Notes), 2024, pp. 2424–2431.

[11] 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.

[12] A. Hashemi, W. Shi, A survey on writing style change detection: Current literature and future directions, Machine Intelligence Research 22 (2025) 397–416.

[13] V. A. Oloo, C. Otieno, L. A. Wanzare, A literature survey on writing style change detection based on machine learning: State-of-the-art-review, Int J Comput Trends Technol 70 (2022) 15–32.

[14] A. Hashemi, W. Shi, Enhancing writing style change detection using transformer-based models and data augmentation, in: CLEF (Working Notes), 2023.

[15] T.-M. Lin, C.-Y. Chen, Y.-W. Tzeng, L.-H. Lee, Ensemble pre-trained transformer models for writing style change detection., in: CLEF (Working Notes), 2022, pp. 2565–2573.

[16] A. A. Khan, M. Rai, K. A. Khan, S. J. Shah, F. Alvi, A. Samad, Team gladiators at pan: Improving author identification: A comparative analysis of pre-trained transformers for multi-author classification., in: CLEF (Working Notes), 2024, pp. 2688–2694.

[17] M. K. Sheykhlan, S. K. Abdoljabbar, M. N. Mahmoudabad, Team karami-sh at pan: Transformer-based ensemble learning for multi-author writing style analysis., in: CLEF (Working Notes), 2024, pp. 2676–2681.

[18] X. Jiang, H. Qi, Z. Zhang, M. Huang, Style change detection: Method based on pre-trained model and similarity recognition., in: CLEF (Working Notes), 2022, pp. 2526–2531.

[19] J. Lv, Y. Yi, H. Qi, Team fosu-stu at pan: Supervised fine-tuning of large language models for multi author writing style analysis., in: CLEF (Working Notes), 2024, pp. 2781–2786.

[20] I. Boriceanu, A. Băltoiu, Style change detection using graph and structural-linguistic features for multi-author writing analysis, in: Working Notes of CLEF 2025, 2025.

[21] 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: 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.