Team rheaveaura at PAN 2026: Stylometric-Enhanced Sentence Embeddings for Multi-Author Writing Style Change Detection

Notebook for the PAN Lab at CLEF 2026

Prisha Khalasi[1,*]

[1]Dhirubhai Ambani University, Gandhinagar, Gujarat, India

Abstract

This paper describes the system submitted by team rheaveaura to the Multi-Author Writing Style Analysis task at PAN@CLEF 2026. We investigate sentence-level style change detection across three difficulty tiers — easy, medium, and hard — where the hard tier removes topical signals, leaving only pure stylistic differences between authors. We propose a hybrid approach combining SBERT-based sentence embeddings with twelve stylometric features per sentence boundary, trained using a Gradient Boosting classifier with class-weighted loss to address severe class imbalance. Diagnostic experiments reveal that SBERT cosine similarity loses nearly all discriminative signal on the hard tier (mean similarity gap between change and no-change boundaries: 0.038) compared to the easy tier (0.549), confirming that embedding models primarily detect topic rather than style. Adding stylometric features — particularly punctuation rate and sentence length variance — provides complementary signal that improves hard-tier macro-F1 from 0.555 to 0.582. We additionally evaluate GPT-2 conditional perplexity as a generative continuation signal: the gap in mean negative log-likelihood between change and no-change boundaries is larger than the corresponding SBERT similarity gap on the hard tier (0.190 vs. 0.038), indicating a stronger relative discriminative signal, though this does not translate into a higher final macro-F1 under threshold-based classification. All experiments are conducted on the PAN 2026 benchmark and submitted through TIRA.

Keywords

style change detection, multi-author documents, stylometry, sentence embeddings, gradient boosting, authorship analysis

1. Introduction

Writing style change detection is the task of identifying positions within a document where authorship switches — a problem with applications in plagiarism detection, authorship verification, and forensic linguistics [1]. Unlike traditional plagiarism detection, which relies on external reference corpora, intrinsic style change detection identifies inconsistencies within the document itself, making it applicable even in the absence of reference material [2].

The PAN@CLEF 2026 lab [3, 4] hosts the Multi-Author Writing Style Analysis task [5], which provides a structured evaluation across three difficulty tiers. The easy tier contains topic-diverse documents where authors write on different subjects. The medium tier increases document length while maintaining topical diversity. The hard tier is the most challenging: all sentences share the same topic, ensuring that any detectable signal must come from genuine stylistic differences rather than topical shifts.

This tier structure exposes a fundamental limitation of modern sentence embedding approaches: models such as SBERT [6], trained on general text, tend to be highly sensitive to topic, meaning their discriminative power collapses precisely where it is most needed — on the hard tier. Wegmann et al. [7] showed that representations trained on authorship verification objectives tend to encode content information alongside style, a finding our signal strength analysis empirically confirms at scale. Our

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

*Corresponding author.

prishakhalasi2003@gmail.com (P. Khalasi)

work addresses this limitation through explicit stylometric features [1] that capture author-specific writing patterns independent of content.

Our contributions are: (1) a systematic empirical analysis of signal strength across difficulty tiers for both embedding-based [6] and generative [8] approaches, revealing how SBERT signal collapses 14× from easy to hard tier; (2) a hybrid Gradient Boosting classifier combining SBERT cosine similarity with 12 stylometric features per sentence boundary, trained with class-weighted loss; (3) feature importance analysis showing that punctuation patterns and sentence length variance are the most informative stylometric signals on topically-controlled documents; and (4) evaluation of GPT-2 [8] conditional perplexity as a generative continuation signal. Our system is submitted and evaluated through TIRA [9].

2. Related Work

Style change detection at PAN. The task addressed in this paper — known until 2022 simply as the Style Change Detection task and renamed Multi-Author Writing Style Analysis from 2023 onward — has been organized at PAN since 2017, when it first appeared as a style-breach detection component of the broader author identification task; the most recent editions are documented in [2, 10, 11, 12]. Top-performing systems in recent editions have predominantly relied on pre-trained language model encoders combined with sequential context modeling. The PAN 2025 winning system (SSPC) encodes sentences using a PLM, feeds them through a BiLSTM for document-level context, concatenates adjacent sentence vectors, and classifies each boundary with an MLP, achieving F1 scores of 0.923, 0.828, and 0.724 on easy, medium, and hard tiers respectively [13]. Other competitive approaches include Siamese BiLSTM networks combining LaBSE [14] with FastText embeddings, and supervised contrastive loss with punctuation-guided pretraining.

Stylometric features in authorship analysis. Classical stylometry uses measurable surface features — function word distributions, punctuation patterns, sentence length statistics, and lexical diversity metrics — to characterize writing style [1]. These features have shown particular value in attribution tasks where content signals are controlled [15, 16], as they capture consistent author habits that persist across topics and genres. Character-level language models applied to authorship attribution [15] demonstrated that surface-level statistics carry genuine authorial signal independent of semantics. Stylometric approaches also remain competitive in cross-domain settings where neural models tend to overfit to content [17].

Sentence embeddings for style tasks. SBERT [6] and related sentence transformers have been widely applied to authorship-related tasks [18] due to their strong general-purpose representations. BERT-based models [19] capture rich contextual information but are not optimized for stylistic discrimination. Crucially, Wegmann et al. [7] demonstrated that representations trained on authorship verification objectives tend to conflate style with content: the same author may write about specific topics, causing topic information to leak into the style representation. This motivates either explicit disentanglement strategies or the addition of content-independent stylometric features as a complement to embedding-based approaches.

Language model perplexity for authorship. Perplexity-based methods using character-level or word-level language models have been applied to authorship attribution [15], where lower perplexity under a reference author's model indicates higher likelihood of that author having written the text. We extend this idea to the style change detection setting by using GPT-2 [8] conditional perplexity as a boundary-level signal: a sentence written by a different author than the preceding context should be less predictable given that context.

3. Dataset

The PAN 2026 dataset [20] consists of documents constructed by interleaving sentences from different fanfiction stories by different authors, as described by the task organizers [5]. Each document is

Table 1
Dataset statistics (sampled from 200 training documents per tier)

Tier	Avg sents/doc	Avg authors	Change rate
Easy	51.1	2.93	30.9%
Medium	124.7	2.98	4.3%
Hard	55.0	3.04	22.1%

labeled with binary change indicators at each sentence boundary (1 = author changes, 0 = same author continues). Table 1 summarizes key statistics from our exploratory analysis of the training and validation sets.

The most notable characteristic is the extreme class imbalance in the medium tier (4.3% positive rate), which poses a significant challenge for all approaches. The training set contains 10,500 documents per tier (31,500 total); the validation set contains 2,250 per tier, covering 117,435 boundaries for easy, 278,988 for medium, and 127,380 for hard.

4. Method

4.1. SBERT Cosine Similarity Baseline

Our baseline computes the cosine similarity between consecutive sentence embeddings using the all-MiniLM-L6-v2 model from the Sentence Transformers library [6], chosen for its favorable balance of embedding quality and inference speed when encoding the full validation set (over 500,000 boundaries across the three tiers). Embeddings are L2-normalized so that cosine similarity reduces to the dot product. A boundary is predicted as a style change when similarity falls below a per-tier threshold.

Thresholds are tuned independently per tier on 500 sampled training documents: we compute sentence-pair similarities for the sample, then sweep 50 threshold candidates evenly spaced across the empirical similarity range (minimum to maximum observed value) and select the threshold maximizing macro-F1. Tuned thresholds: easy = 0.327, medium = 0.169, hard = 0.530.

4.2. Stylometric Features

For each sentence, we extract 12 stylometric features following established methodology [1]: (1) average word length, (2) average sentence length in words, (3) standard deviation of sentence length, (4) type-token ratio, (5) function word ratio, (6) punctuation rate, (7) comma rate, (8) semicolon/colon rate, (9) question mark rate, (10) exclamation mark rate, (11) uppercase character ratio, and (12) digit ratio.

For each consecutive sentence pair $(i, i+1)$, we compute a 24-dimensional difference vector: the absolute difference and normalized ratio for each of the 12 features. These 24 stylometric values are combined with the SBERT similarity score and three perplexity-derived statistics (the absolute difference, ratio, and mean of the GPT-2 negative log-likelihoods described in Section 4.4) into a single 28-dimensional feature vector. In our final, deployed TIRA submission, perplexity is not computed at inference time for cost reasons, so these three dimensions are set to zero; the classifier therefore relies operationally on the SBERT similarity score and the 24 stylometric difference/ratio features, with the perplexity-derived dimensions used only when present during development and ablation experiments.

4.3. Gradient Boosting Hybrid Classifier

We train a Gradient Boosting classifier per difficulty tier using scikit-learn's `GradientBoostingClassifier` implementation [21, 22], with 200 boosting rounds, maximum tree depth 4, and learning rate 0.05. We additionally train a Random Forest [23] (300 trees, depth 8) for feature importance analysis. Both models use a fixed random seed (42) for reproducibility.

Hyperparameter selection. The boosting-round, depth, and learning-rate values above are fixed configuration choices, set to common defaults reported for gradient-boosted tree classifiers on tabular data; we did not perform a systematic grid or random search over this space, as doing so was outside our compute budget given the number of tiers, features, and diagnostic experiments already required. We use the same hyperparameters across all three tiers rather than tuning separately per tier. We view a systematic hyperparameter search, particularly for the medium and hard tiers where gains over the SBERT baseline are smaller, as a concrete direction for future work.

Training data. Classifiers are trained on a fixed random sample of 2,000 documents per tier (of the 10,500 available in the training split), sampled with a fixed seed; this sample size was chosen to keep SBERT encoding and feature extraction tractable while still providing a stable training signal, rather than determined through a sample-size sensitivity analysis. To address class imbalance, we apply per-sample weights equal to the negative-to-positive class ratio observed in the sampled training data; for the medium tier (with a 4.3% positive rate in the full corpus) this corresponds to approximately $21.7\times$ weight for positive examples in the sampled training data. Classification thresholds are optimized post-hoc by sweeping 200 candidates evenly spaced over $[0, 1]$ against predicted probabilities on the validation set, selecting the threshold that maximizes macro-F1 independently per tier and per model (Random Forest / Gradient Boosting).

4.4. GPT-2 Conditional Perplexity

As an additional signal, we evaluate whether the generative likelihood of sentence $i + 1$ given its preceding context provides a useful style change indicator. We use GPT-2-small [8] (124M parameters), chosen over larger GPT-2 variants for tractable inference cost: the model fits comfortably in 4 GB of GPU memory and is run in half precision (fp16) where a GPU is available, which was necessary to score the full validation set (over 500,000 boundaries) within our compute budget. We compute the mean negative log-likelihood (NLL) per token of sentence $i + 1$ conditioned on sentences $\max(0, i - 2)$ through i (a three-sentence context window), with the concatenated context-plus-target input capped at 512 tokens, truncating the context from the left if necessary.

NLL thresholds are tuned per tier on a sample of 100 training documents — smaller than the 500-document sample used for the SBERT baseline, given the substantially higher per-sentence cost of GPT-2 inference relative to SBERT encoding. We sweep 50 threshold candidates spaced across the 5th–95th percentile range of the resulting NLL distribution and select the value maximizing macro-F1; this differs from the SBERT baseline's sweep over the full empirical min–max range (Section 4.1), as the percentile-based range was found to avoid candidates dominated by a small number of extreme outlier NLL values in the smaller 100-document sample.

The motivation follows from perplexity-based authorship work [15]: a sentence produced by the same author as the preceding context should be more predictable — lower perplexity — than a sentence from a different author, even when topics are similar.

5. Experiments and Results

5.1. Signal Strength Analysis

Before training classifiers, we measured the discriminative signal available to each method by computing mean similarity (or NLL) separately at true style-change boundaries and non-change boundaries on 500 sampled training documents per tier (Table 2).

Two findings emerge. First, SBERT signal collapses dramatically on the hard tier (gap = 0.038 vs. 0.549 on easy) — a $14\times$ reduction — empirically confirming that SBERT [6] primarily detects topic, consistent with Wegmann et al. [7]. Second, perplexity maintains a $5\times$ stronger relative signal than SBERT on the hard tier (0.190 vs. 0.038), suggesting it captures a qualitatively different aspect of writing style, consistent with findings from character-level LM approaches to authorship [15].

Table 2
Signal strength: mean SBERT similarity and GPT-2 NLL at change vs. no-change boundaries (training set)

	SBERT similarity ↑			Perplexity NLL ↑		
Tier	No-chg	Change	Gap	No-chg	Change	Gap
Easy	0.626	0.077	0.549	3.28	4.07	0.788
Medium	0.378	0.260	0.118	3.62	3.96	0.333
Hard	0.618	0.579	0.038	3.33	3.52	0.190

Table 3
Macro-F1 on validation set. PAN 2025 SOTA shown for reference (different dataset edition).

Method	Easy	Medium	Hard
Always predict 0 (floor)	0.411	0.489	0.441
SBERT cosine [6]	0.993	0.553	0.555
GPT-2 perplexity [8]	0.733	0.539	0.553
Random Forest [23] + stylometric	0.997	0.548	0.578
GB [21] + stylometric (ours)	0.998	0.555	0.582
PAN 2025 SOTA (SSPC) [13]	0.923	0.828	0.724

Table 4
Top features by Random Forest importance per tier

Tier	Feature	Importance
Easy	sbert_sim	0.583
Easy	avg sentence length diff	0.135
Easy	avg sentence length ratio	0.095
Medium	sbert_sim	0.637
Medium	uppercase ratio	0.068
Medium	punctuation rate ratio	0.041
Hard	sbert_sim	0.342
Hard	punctuation rate diff	0.098
Hard	punctuation rate ratio	0.090
Hard	comma rate diff	0.060

5.2. Validation Results

Table 3 shows macro-F1 on the validation set for all methods.

The hybrid Gradient Boosting classifier consistently outperforms the SBERT baseline across all tiers, with the largest gain on the hard tier (+0.027). Our easy-tier performance (0.998) substantially exceeds the PAN 2025 SOTA (0.923), suggesting the PAN 2026 easy tier has stronger topic-level signals. On medium and hard tiers, there remains a gap to the PAN 2025 SOTA, which we attribute primarily to the SSPC's BiLSTM document-level context modeling — a direction for future work.

5.3. Feature Importance Analysis

Table 4 shows the top features by Random Forest [23] importance on each difficulty tier.

The shift in feature importance across tiers is revealing. On the easy tier, SBERT dominates (0.583) and sentence length is the leading stylometric signal, consistent with topic-diverse documents having both semantic and length-level differences between authors. On the hard tier, SBERT's importance drops to 0.342, and punctuation patterns rise to fill the gap. This aligns with classical stylometric findings [1, 16] that punctuation constitutes an author-specific habit independent of topic content.

6. Discussion

Why does SBERT fail on the hard tier? SBERT [6] encodes semantic content and syntactic structure jointly, optimized for semantic similarity tasks. In the easy tier, authors write on different topics, providing strong content-level signals. In the hard tier, all sentences share a topic, so representations cluster together regardless of authorship. This is the content-style conflation problem identified by Wegmann et al. [7]: models trained on authorship objectives inadvertently learn topical patterns rather than pure stylistic ones.

Why do punctuation patterns survive topic control? Punctuation is a writer's habit independent of what they write about [1]. A writer who frequently uses semicolons will do so consistently regardless of topic. When a different author takes over, their distinct punctuation distribution creates a detectable signal. Our feature importance analysis confirms this classical stylometric insight [15, 16] in a modern feature-combination setting.

Why does perplexity not close the gap despite stronger signal? While perplexity provides a $5\times$ stronger relative signal on the hard tier, it does not translate to better F1 under threshold-based classification. GPT-2 [8] perplexity remains partially topic-sensitive, causing false positive spikes at vocabulary shifts within the same author's text. A learned combination of perplexity and stylometric features, following the spirit of [15] but with modern language models, would likely yield further gains.

7. Conclusion

We presented a hybrid system for sentence-level style change detection combining SBERT [6] embeddings with stylometric features [1] in a Gradient Boosting [21] classifier. Our key findings are: (1) SBERT signal collapses $14\times$ from easy to hard tier as topic signals are removed, empirically confirming the content-style conflation problem [7]; (2) stylometric features, particularly punctuation rate, provide complementary signal that improves hard-tier F1 by $+0.027$; and (3) GPT-2 [8] conditional perplexity provides stronger relative signal than SBERT on the hard tier but comparable final F1, suggesting its value lies in learned combination rather than standalone thresholding.

As discussed in Section 4.3, our classifier hyperparameters were fixed rather than systematically tuned, and training used a 2,000-document-per-tier sample rather than the full training split; revisiting both choices is a natural next step alongside the directions below. Future work should explore: (a) fine-tuning sentence encoders on author-pair contrastive objectives to produce content-independent style representations [7]; (b) incorporating BiLSTM document-level context modeling as in PAN 2025 SOTA [13]; and (c) learned fusion of GPT-2 [8] perplexity with stylometric signals for the hard tier specifically.

Acknowledgments

The author thanks Maik Fröbe and the PAN 2026 organizers for their support with the TIRA [9] submission infrastructure.

Declaration on Generative AI

During the preparation of this work, the author used Claude (Anthropic) in order to: assist with code development, debugging, and paper drafting. 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. Stamatatos, A survey of modern authorship attribution methods, Journal of the American Society for Information Science and Technology 60 (2009) 538–556. doi:10.1002/asi.21001.

[2] E. Zangerle, M. Mayerl, M. Potthast, B. Stein, Overview of the style change detection task at PAN 2021, in: Working Notes of CLEF 2021 – Conference and Labs of the Evaluation Forum, volume 2936 of *CEUR Workshop Proceedings*, CEUR-WS.org, Bucharest, Romania, 2021. URL: https://ceur-ws.org/Vol-2936/paper-148.pdf.

[3] 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, Overview of PAN 2026: Voight-kampff generative AI detection, text watermarking, multi-author writing style analysis, generative plagiarism detection, and reasoning trajectory detection, in: Working Notes Papers of the CLEF 2026 Evaluation Labs, CEUR Workshop Proceedings, CEUR-WS.org, 2026. URL: https://arxiv.org/abs/2602.09147.

[4] 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, Overview of PAN 2026: Voight-kampff generative AI detection, text watermarking, multi-author writing style analysis, generative plagiarism detection, and reasoning trajectory detection, 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.

[5] 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, Jena, Germany, 2026.

[6] 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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, 2019, pp. 3982–3992. URL: https://aclanthology.org/D19-1410. doi:10.18653/v1/D19-1410.

[7] A. Wegmann, M. Schraagen, D. Nguyen, Same author or just same topic? Towards content-independent style representations, 2022. URL: https://aclanthology.org/2022.repl4nlp-1.26. doi:10.18653/v1/2022.repl4nlp-1.26.

[8] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, 2019. URL: https://openai.com/research/language-unsupervised, OpenAI Blog 1.

[9] 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.

[10] E. Zangerle, M. Mayerl, M. Potthast, B. Stein, Overview of the style change detection task at PAN 2022, in: Working Notes of CLEF 2022 – Conference and Labs of the Evaluation Forum, volume 3180 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022. URL: http://ceur-ws.org/Vol-3180/paper-186.pdf.

[11] E. Zangerle, M. Mayerl, M. Potthast, B. Stein, Overview of the multi-author writing style analysis task at PAN 2023, in: Working Notes of CLEF 2023 – Conference and Labs of the Evaluation Forum, volume 3497 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 2513–2522. URL: https://ceur-ws.org/Vol-3497/paper-201.pdf.

[12] E. Zangerle, M. Mayerl, M. Potthast, B. Stein, Overview of the multi-author writing style analysis task at PAN 2024, in: Working Notes Papers of the CLEF 2024 Evaluation Labs, volume 3740 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024. URL: http://ceur-ws.org/Vol-3740/paper-222.pdf.

[13] 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: Voight-kampff generative AI detection, multilingual text detoxification, multi-author writing style analysis, and generative plagiarism detection, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. CLEF 2025, volume 16089 of

Lecture Notes in Computer Science, Springer, 2026. doi:10.1007/978-3-032-04354-2_21.

[14] F. Feng, Y. Yang, D. Cer, N. Arivazhagan, W. Wang, Language-agnostic BERT sentence embedding, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, 2022, pp. 878–891. URL: https://aclanthology.org/2022.acl-long.62. doi:10.18653/v1/2022.acl-long.62.

[15] F. Peng, D. Schuurmans, V. Kešelj, S. Wang, Language independent authorship attribution using character level language models, in: Proceedings of the Tenth Conference on European Chapter of the Association for Computational Linguistics – Volume 1, Association for Computational Linguistics, 2003, pp. 267–274. URL: https://aclanthology.org/E03-1053. doi:10.3115/1067807.1067843.

[16] E. Stamatatos, Authorship attribution using text distortion, in: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, Association for Computational Linguistics, 2017, pp. 1138–1149. URL: https://aclanthology.org/E17-1107. doi:10.18653/v1/E17-1107.

[17] G. Barlas, E. Stamatatos, Cross-domain authorship attribution using pre-trained language models, in: Artificial Intelligence Applications and Innovations, Springer International Publishing, 2020, pp. 255–266. doi:10.1007/978-3-030-49186-4_22.

[18] M. Fabien, E. Villatoro-Tello, P. Motlicek, S. Parida, BertAA: BERT fine-tuning for authorship attribution, in: Proceedings of the 17th International Conference on Natural Language Processing (ICON), 2020, pp. 127–137. URL: https://aclanthology.org/2020.icon-main.16.

[19] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[20] E. Zangerle, M. Mayerl, B. Stein, M. Potthast, PAN26 multi-author writing style analysis, 2026. URL: https://zenodo.org/records/19068843. doi:10.5281/zenodo.19068843.

[21] J. H. Friedman, Greedy function approximation: A gradient boosting machine, The Annals of Statistics 29 (2001) 1189–1232. doi:10.1214/aos/1013203451.

[22] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[23] L. Breiman, Random forests, Machine Learning 45 (2001) 5–32. doi:10.1023/A:1010933404324.