Team ULILLE at PAN 2026: Task-Routed Ettin Cross-Encoders for Multi-Author Writing Style Analysis

Notebook for the PAN Lab at CLEF 2026

Abderaouf Khelfaoui[1], Damien Sileo[2] and Gabriel Loiseau[2,3]

[1]University of Lille, Lille, France
[2]University of Lille, Inria, CNRS, Centrale Lille, UMR 9189 - CRIStAL, F-59000 Lille, France
[3]Hornetsecurity, Hem, France

Abstract

This paper addresses the PAN 2026 Multi-Author Writing Style Analysis task, which requires detecting the sentence-level positions at which the author changes in a document. The 2026 edition introduces three evaluation regimes – easy, medium, and hard – that progressively constrain how much topical evidence accompanies the authorial signal, ranging from documents with free topical variation to documents whose sentences all share a single topic. Instead of training a single classifier to absorb all three regimes, we cast the problem as binary classification over adjacent sentence pairs and fine-tune one Ettin encoder cross-encoder per regime. Our main contribution is a document-level task router that infers which regime an unseen document belongs to and dispatches it to the corresponding specialist, yielding a modular system in which each component is optimised for a well-defined slice of the writing style change detection problem.

Keywords

Style change detection, Multi-author writing analysis, Authorship analysis, Cross-encoder, Ettin, PAN 2026

1. Introduction

Multi-author writing style analysis is the task of identifying whether and where a text contains stylistic discontinuities that indicate a change of author. The PAN shared tasks have studied this problem for several years, gradually moving from document- and paragraph-level formulations toward fine-grained sentence-level detection [1]. In PAN 2026, each document is represented as a sequence of sentences, and participants must output one binary decision for each adjacent sentence pair: 1 if the author changes between the two sentences and 0 otherwise [2].

The 2026 task provides three regimes that vary in how much topical evidence accompanies the authorial signal [3]. In the easy setting, sentences in a document cover a variety of topics, so a model can in principle exploit topic shifts as a proxy for authorship boundaries. In the medium setting, the topical variety is reduced, weakening this proxy. In the hard setting, all sentences in a document are on the same topic, which removes topical evidence by construction and forces the model to rely on style rather than content. Recent authorship analysis work has emphasized that disentangling content from style is still an open problem: even strong representations such as universal authorship embeddings [4] and style-targeted models [5] can confound the two signals, and cross-genre benchmarks [6] show that authorship classifiers degrade substantially when topical statistics shift. Read in this light, we interpret the easy/medium/hard split as three distinct points along the content–style axis: the easy regime is closer to topic-aided authorship analysis, the hard regime is closer to pure style change detection, and the medium regime sits in between with weak topical evidence. A single global decision rule is consequently not equally well suited to all three regimes, since both the informative cues and the positive boundary rate vary substantially across tracks.

Our submission therefore uses a task-routed specialist architecture rather than a single global classifier. We train one sentence-pair cross-encoder per regime so that each specialist can settle on a decision

CLEF 2026 Working Notes, September 2026
abderaouf.khelfaoui.etu@univ-lille.fr (A. Khelfaoui); damien.sileo@inria.fr (D. Sileo); gabriel.loiseau@inria.fr (G. Loiseau)

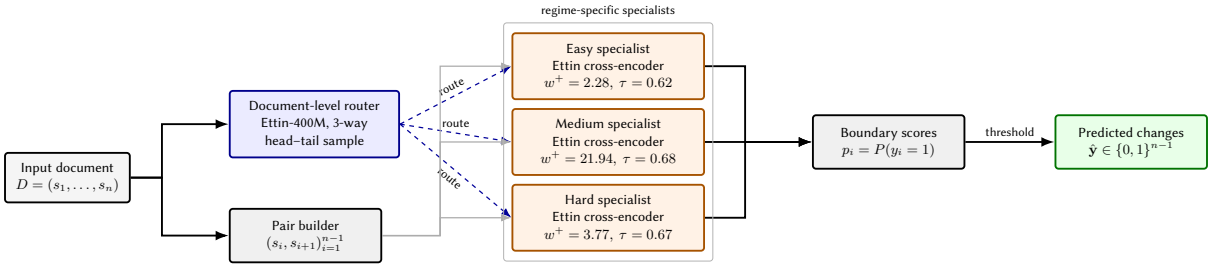

Figure 1: Task-routed inference pipeline. A document-level Ettin router assigns each document to one regime, while adjacent sentence pairs are scored by the corresponding regime-specific Ettin cross-encoder. Scores are converted to sentence-boundary decisions using validation-tuned thresholds.

surface, loss weighting, and threshold matched to its own content–style mix, and we train a separate document-level task router that predicts which regime an incoming document belongs to. Our system loads the router and the three specialists, predicts a route for each document, applies the selected specialist to all adjacent sentence pairs, and thresholds the resulting scores.

2. Related Work

Style change detection belongs to intrinsic authorship analysis: the system must identify stylistic discontinuities inside a document without access to reference texts by the candidate authors. This distinguishes it from closed-set authorship attribution and makes the problem close to authorship verification, where similar passages have to be found from internal evidence alone. Classical authorship analysis has long used lexical, character-level, syntactic, and structural regularities as stylometric evidence [7]; PAN style-change tasks adapt this idea from assigning a whole document to an author toward locating the boundaries at which the authorial signal changes.

Earlier PAN systems relied mainly on handcrafted stylometric features and shallow models, including text statistics, character-level CNNs, ensembles of stylistic cues, clustering approaches, and neural sentence embeddings [8, 9, 10, 11]. These works established the core intuition of style change detection: author shifts can be identified through local stylistic patterns such as punctuation, sentence structure, and function-word usage rather than topic alone.

More recent PAN editions shifted toward pretrained transformers and local transition modeling. PAN 2020 introduced BERT-based detectors, while later systems explored Siamese architectures, similarity learning, BiLSTM/BERT hybrids, contrastive embeddings, and transition-based transformer models [12, 13, 14, 15, 16]. Data augmentation and fine-tuned transformer ensembles became increasingly common [17, 18].

The closest predecessors are the 2025 sentence-level systems, which used transformer ensembles, contrastive DeBERTa models, structural linguistic features, and sequential sentence-pair classifiers [19, 20, 21, 22]. Particularly related to our approach, Team HHU employed multiple specialists for different difficulty regimes within a single architecture [23]. Our work extends this idea by separating content and style effects on multiple models: a document-level router assigns each input to one of three calibrated Ettin cross-encoder specialists based on its content–style regime.

A recurring concern across these editions, and across modern authorship analysis more broadly, is that what looks like an authorial signal can in fact be a topical one. Stylometric work has long recognized that lexical and content features dominate similarity unless they are explicitly controlled [7], and recent representation-learning approaches confirm this: universal authorship embeddings trained on large amounts of social-media text capture both authorial and topical regularities [4], and content-controlled contrastive training was introduced explicitly to push representations toward style and away from topic [5]. Cross-genre and cross-topic benchmarks such as CrossNews [6] report substantial drops when authorship classifiers move across topical or genre boundaries, again pointing to a residual entanglement between content and style. Topic control has also been a recurring axis in PAN's own

Table 1
Training and validation statistics for the three sentence-pair specialist models.

Difficulty	Train pairs	Train positive rate	Validation pairs	Validation positive rate
Easy	549,804	0.3052	117,435	0.3029
Medium	1,299,324	0.0436	278,988	0.0438
Hard	589,946	0.2094	127,380	0.2125

evolution, as summarized in the PAN 2023 overview [24]. The PAN 2026 easy/medium/hard split formalizes this concern at the dataset level: the three regimes ask for the same output but afford different mixtures of content and style cues, so the strength of any topic-based shortcut differs accordingly. Our task-routed design is a direct response to this state of the literature, rather than asking one classifier to be simultaneously topic-aware (where helpful) and topic-invariant (where required), we let each specialist commit to one operating regime and let the router decide which regime applies.

Our pairwise classifier is also related to bi-encoder sentence-similarity approaches such as Sentence-BERT [25], where each segment is embedded independently and compared afterward. A cross-encoder is slower, because it recomputes attention jointly for each adjacent pair, but it can model token-level interactions across the candidate boundary. This is useful when a boundary is signaled by subtle shifts in punctuation, discourse markers, syntactic choices, or lexical preferences rather than by an obvious topical jump.

3. Task and Data

Each dataset instance consists of a document already divided into sentences. The goal is to determine whether the author changes between consecutive sentence pairs. Performance is measured with macro-F1, computed independently for the three difficulty regimes.

Table 1 summarizes the data used by our final specialists and router. The validation positive rate highlights the value of regime-specific modeling. The easy set has many positive boundaries, the hard set has a moderate positive rate, and the medium set contains only about 4.4% positive boundaries. A threshold or loss weighting strategy that works well for easy is therefore not necessarily appropriate for medium.

The router is trained on whole documents rather than sentence pairs. It uses 31,500 training documents and 6,750 validation documents, balanced across the three regime labels. During preprocessing, each document is represented by a head-tail sample of up to 64 sentences and tokenized to a maximum length of 4096 tokens.

We did not use external text corpora or author metadata. The only supervision used for the specialists is the official sentence-boundary label sequence, and the only supervision used for the router is the regime label implied by the official split. This makes the system simple to reproduce, but it also means that the router learns dataset-level regularities of the data construction procedure rather than a content–style classifier in a general sense.

4. Method

4.1. Sentence-Pair Formulation

Given a document with sentences $s_1, \ldots, s_n$, we model the task as predicting whether an author change occurs between each consecutive sentence pair (s_i, s_{i+1}). This pairwise formulation matches the official sentence-boundary evaluation setup and avoids additional post-processing from larger text segments back to sentence-level decisions.

We considered this formulation a simple and reliable baseline for the task. Although it does not explicitly model longer-range discourse context, it has two practical advantages. First, each prediction

Table 2
Validation confusion matrix for the task router. Rows are true regime labels and columns are predicted labels.

	Pred. easy	Pred. medium	Pred. hard
True easy	2,195	0	55
True medium	0	2,249	1
True hard	33	1	2,216

corresponds directly to one required output label. Second, it exposes class imbalance at the boundary level, making it easier to tune loss weights and decision thresholds separately for each regime.

4.2. Specialist Cross-Encoders

Each specialist is a cross-encoder initialized from `jhu-clsp/ettin-encoder-400m`[1]. Ettin is an open suite of paired encoder and decoder models trained for controlled comparisons across sequence-modeling objectives [26]; the released encoder model used here is exposed through the Hugging Face Transformers sequence-classification interface and tagged as a ModernBERT-family model [27, 28]. We use it as a classification backbone over adjacent sentence pairs. The cross-encoder receives the two sentences jointly, allowing self-attention to compare lexical, syntactic, topical, and stylistic cues across the boundary.

We train three independent specialists, one for each regime. Each model is optimized with binary cross-entropy. To account for the strong class imbalance, especially in medium, we set the positive-class weight automatically to the ratio of negative to positive training pairs for that regime. The resulting positive weights are 2.28 for easy, 21.94 for medium, and 3.77 for hard. All specialists are trained for one epoch with learning rate 2×10^{-5}, batch size 32, maximum sequence length 256, weight decay 0.01. After training, we sweep thresholds from 0.1 to 0.9 on the validation set and keep the threshold that maximizes macro-F1 for each specialist.

The maximum length of 256 tokens was sufficient for the sentence-pair inputs in this dataset while keeping inference manageable. We did not tune the backbone architecture or train multi-seed ensembles. The development effort instead focused on matching the classifier, loss weighting, and decision threshold to each official regime. We trained all models with Nvidia A30 GPU card with 24GB memory and Intel Xeon Gold 5320 CPU.

4.3. Task Router

The final system cannot assume that the difficulty (i.e. task) is available or reliable in every execution layout, but more importantly the three official regimes can be seen as different authorship analysis tasks rather than three reproductions of one task. Easy documents allow the classifier to use topical discontinuities as a boundary cue; hard documents remove that cue by construction, so any boundary signal must come from style; medium documents fall in between, with weak topical evidence and a much lower positive boundary rate. A single classifier trained jointly over all three regimes would have to reconcile incompatible feature priorities and label priors, and would in practice learn an averaged operating point that is suboptimal for each regime. We therefore train a document-level task router, also initialized from `ettin-encoder-400m`, as a three-way classifier over easy, medium, and hard. The router uses the first and last sentences of the document, capped at 64 sentences total, enabling to capture document-level topic distribution, topical coherence, and length patterns that are useful indicators of which content–style regime a document belongs to.

On validation, the router reaches 0.9867 accuracy. The main confusions are between easy and hard; medium is almost perfectly separated. The confusion matrix is shown in Table 2. Routing medium correctly is especially important because medium has a very different boundary prevalence and uses a substantially different specialist threshold and class weight.

[1] https://hf.co/jhu-clsp/ettin-encoder-400m

Table 3
Macro-F1 scores on validation and hidden test data. The average is the unweighted mean over the three regimes and is included only as a compact summary; PAN evaluates the three tracks independently.

Difficulty	Validation macro-F1	Test macro-F1
Easy	0.9998	0.995
Medium	0.7179	0.712
Hard	0.7657	0.769
Average	0.8278	0.8253

Table 4
Internal validation comparison between an earlier shared cross-encoder and the final routed specialists.

Model	Easy	Medium	Hard	Mean
Shared Ettin-400M cross-encoder	0.8822	0.6143	0.6427	0.7131
Routed Ettin-400M specialists	0.9998	0.7179	0.7657	0.8278

4.4. Inference

At inference time, our system predicts a task route for each document, groups documents by route, and loads each specialist only for the documents routed to it. This grouping avoids keeping all specialists active for every batch, which was useful for memory management. For every adjacent sentence pair, the selected specialist outputs a score. The score is converted to a binary boundary decision using the validation-tuned threshold of that specialist: 0.62 for easy, 0.68 for medium, and 0.67 for hard. In practice, the only learned artifacts required at test time are the router checkpoint, the three specialist checkpoints, their tokenizer files, and a small configuration file containing the per-track thresholds.

5. Results

Table 3 reports validation and hidden test macro-F1 scores. The validation scores are measured on the official validation split using the same per-regime thresholds later embedded in the submitted software. The test scores are the official scores obtained through the TIRA evaluation platform [29] after the shared task.

The validation-to-test behavior is stable. Easy remains close to perfect, decreasing from 0.9998 to 0.995. Medium decreases slightly from 0.7179 to 0.712, which is consistent with its low positive rate and the high cost of false positives under macro-F1. Hard slightly improves from 0.7657 to 0.769. The resulting average changes by less than 0.003 points.

Table 4 places the final specialists in context with an earlier shared Ettin-400M cross-encoder development run with the same hyperparameters. The shared model was useful for validating the sentence-pair framing, but the routed specialist configuration substantially improves all three tracks, especially easy and medium. Because the comparison changes both model capacity and specialization.

6. Discussion

The main benefit of the task-routed specialist approach is that it leverages the different content–style regimes rather than treating them as gradations of a single task. The three regimes differ in two ways at once. They differ in label priors (the medium set has far fewer positive boundaries than easy and hard, so its classifier requires stronger class weighting and careful threshold calibration) and they differ in what kind of authorship signal is actually available. The easy specialist can lean on topic shifts as a partial proxy for boundaries, the hard specialist must work with style alone because topic is held constant by construction, and the medium specialist sits in between with weak topical evidence

and rare boundaries. A shared classifier would be forced into a compromise across these incompatible regimes; separating the models lets each specialist learn a decision surface and operating point matched to its own content–style mix.

The easy score should not be interpreted as evidence of style understanding. In the easy regime, topical discontinuity is itself a strong boundary cue, and modern authorship representations are known to confound topical and authorial regularities [5, 4]. The hard score is lower but transfers well to the hidden test set, indicating that when the topical shortcut is removed by construction the cross-encoder still learns useful non-topical style cues. Taken together, the easy–hard gap quantifies, in our setting, how much of the apparent "authorship signal" comes from topic when topic is allowed to vary.

The router also has an asymmetric risk profile. Misrouting a medium document is potentially costly because the medium specialist uses a much larger positive-class weight and a threshold calibrated for rare boundaries. By contrast, the observed easy/hard confusions are less damaging when the selected specialist still faces a boundary rate in a similar range. This is why the router confusion matrix is more reassuring than its raw accuracy alone suggests: almost all medium documents are routed correctly, so the regime in which content–style separation matters most is the regime that the router handles best.

Several limitations remain. First, the specialists classify each boundary independently, so they cannot enforce document-level consistency such as author segment continuity or plausible alternation patterns. Second, the head-tail router may miss evidence that appears only in the middle of a long document. Third, our threshold sweep is coarse and was tuned on a single validation split; more robust calibration strategies or cross-validation could improve the medium track in particular. Fourth, we did not investigate specialist-specific hyperparameter optimization and instead focused on identifying a single set of strong hyperparameters shared across all models. Future work should therefore explore dedicated hyperparameter searches for each specialist, since the three regimes differ substantially. Finally, the near-perfect easy result leaves limited room for diagnosing how much of the model's behaviour reflects style and how much reflects topic transitions, and our setup does not yet incorporate explicit content-controlled training in the spirit of [5], which would be a natural next step for improving robustness on the hard and medium regimes.

7. Conclusion

We presented a task-routed sentence-pair cross-encoder system for PAN 2026 multi-author writing style analysis. Motivated by the observation that the easy, medium, and hard tracks correspond to qualitatively different authorship analysis tasks along the content–style axis which remains a hard problem in modern authorship representations ; we fine-tune one Ettin-400M specialist per regime, calibrate each specialist on validation macro-F1, and use a document-level Ettin-400M task router to select the specialist at inference time. The submitted solution obtains hidden test macro-F1 scores of 0.995, 0.712, and 0.769 on the easy, medium, and hard tracks, respectively. These results suggest that task-specific calibration and specialist modeling are effective for sentence-level style change detection, while medium-difficulty documents and the broader problem of separating content from style remain the main targets for future improvements.

Declaration on Generative AI

The authors used GPT 5.5 for grammar and spelling checks, then reviewed and edited all content accordingly. Full responsibility for this publication remains with the authors.

References

[1] 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: J. C. de Albornoz, J. Gonzalo, L. Plaza, A. G. S. Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), volume 16089 of *Lecture Notes in Computer Science*, Springer, Berlin Heidelberg New York, 2025, pp. 388–411. doi:10.1007/978-3-032-04354-2_21.

[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] R. A. Rivera-Soto, O. E. Miano, J. Ordonez, B. Y. Chen, A. Khan, M. Bishop, N. Andrews, Learning universal authorship representations, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 913–919. URL: https://aclanthology.org/2021.emnlp-main.70/. doi:10.18653/v1/2021.emnlp-main.70.

[5] A. Wegmann, M. Schraagen, D. Nguyen, Same author or just same topic? towards content-independent style representations, in: Proceedings of the 7th Workshop on Representation Learning for NLP, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 249–268. URL: https://aclanthology.org/2022.repl4nlp-1.26/. doi:10.18653/v1/2022.repl4nlp-1.26.

[6] M. Ma, D. M. Le, J. Kang, Y. Dou, J. Cadigan, D. Freitag, A. Ritter, W. Xu, CrossNews: A cross-genre authorship verification and attribution benchmark, Proceedings of the AAAI Conference on Artificial Intelligence 39 (2025) 24777–24785. URL: https://ojs.aaai.org/index.php/AAAI/article/view/34659. doi:10.1609/aaai.v39i23.34659.

[7] 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.

[8] K. Safin, A. Ogaltsov, Detecting a change of style using text statistics: Notebook for PAN at CLEF 2018, in: Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, volume 2125 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2018. URL: https://ceur-ws.org/Vol-2125/.

[9] N. Schaetti, Character-based convolutional neural network for style change detection: Notebook for PAN at CLEF 2018, in: Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, volume 2125 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2018. URL: https://ceur-ws.org/Vol-2125/.

[10] D. Zlatkova, D. Kopev, K. Mitov, A. Atanasov, M. Hardalov, I. Koychev, P. Nakov, An ensemble-rich multi-aspect approach towards robust style change detection: Notebook for PAN at CLEF 2018, in: Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, volume 2125 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2018. URL: https://ceur-ws.org/Vol-2125/.

[11] K. Safin, R. Kuznetsova, Style breach detection with neural sentence embeddings, in: Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, volume 1866 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2017. URL: https://ceur-ws.org/Vol-1866/.

[12] A. Iyer, S. Vosoughi, Style change detection using BERT, in: Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, volume 2696 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020. URL: https://ceur-ws.org/Vol-2696/.

[13] S. Nath, Style change detection using siamese neural networks, in: Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, volume 2936 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2021, pp. 2073–2082. URL: https://ceur-ws.org/Vol-2936/.

[14] J. Zia, L. Zhou, Z. Liu, Style change detection based on Bi-LSTM and BERT, in: Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, volume 3180 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2022. URL: https://ceur-ws.org/Vol-3180/.

[15] H. Chen, Z. Han, Z. Li, Y. Han, A writing style embedding based on contrastive learning for multi-author writing style analysis, in: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, volume 3497 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 2562–2567. URL: https://ceur-ws.org/Vol-3497/.

[16] I. E. Kucukkaya, U. Sahin, C. Toraman, ARC-NLP at PAN 2023: Transition-focused natural language inference for writing style detection, in: Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, volume 3497 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2023, pp. 2659–2668. URL: https://ceur-ws.org/Vol-3497/.

[17] Y. Huang, L. Kong, Team text understanding and analysis at PAN: Utilizing BERT series pretraining model for multi-author writing style analysis, in: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, volume 3740 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 2653–2657. URL: https://ceur-ws.org/Vol-3740/.

[18] C. Liu, Z. Han, H. Chen, Q. Hu, Team Liuc0757 at PAN: A writing style embedding method based on contrastive learning for multi-author writing style analysis, in: Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, volume 3740 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 2744–2749. URL: https://ceur-ws.org/Vol-3740/.

[19] D. Bölöni-Turgut, D. Verma, C. Cardie, Team cornell-1 at PAN: Ensembling fine-tuned transformer models for writing style analysis, in: CLEF 2025 Working Notes, volume 4038 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025, pp. 3634–3640. URL: https://ceur-ws.org/Vol-4038/.

[20] K. Lin, C. Liu, F. Ye, Z. Han, SCL-DeBERTa: Multi-author writing style change detection enhanced by supervised contrastive learning, in: CLEF 2025 Working Notes, volume 4038 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025, pp. 3781–3787. URL: https://ceur-ws.org/Vol-4038/.

[21] I.-R. Boriceanu, A.-E. Băltoiu, Style change detection using graph and structural linguistic features for multi-author writing analysis, in: CLEF 2025 Working Notes, volume 4038 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025, pp. 3609–3616. URL: https://ceur-ws.org/Vol-4038/.

[22] G. Schmidt, J. Römisch, M. Halchynska, S. Gorovaia, I. P. Yamshchikov, Team "better_call_claude": Style change detection using a sequential sentence pair classifier, in: CLEF 2025 Working Notes, volume 4038 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025, pp. 3916–3928. URL: https://ceur-ws.org/Vol-4038/.

[23] P. Meier, K. Boland, L. Kallmeyer, S. Dietze, Team HHU - an ensemble-based approach to multi-author writing style analysis combining experts for different difficulty levels, in: CLEF 2025 Working Notes, volume 4038 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025, pp. 3842–3849. URL: https://ceur-ws.org/Vol-4038/.

[24] J. Bevendorff, I. Borrego-Obrador, M. Chinea-Ríos, M. Franco-Salvador, M. Fröbe, A. Heini, K. Kredens, M. Mayerl, P. Pezik, M. Potthast, F. Rangel, P. Rosso, E. Stamatatos, B. Stein, M. Wiegmann, M. Wolska, E. Zangerle, Overview of PAN 2023: Authorship verification, multi-author writing style analysis, profiling cryptocurrency influencers, and trigger detection – condensed lab overview, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the 14th International Conference of the CLEF Association (CLEF 2023), Springer, 2023, pp. 459–481. URL: https://doi.org/10.1007/978-3-031-42448-9_29. doi:10.1007/978-3-031-42448-9_29.

[25] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, 2019. URL: https://arxiv.org/abs/1908.10084. arXiv:1908.10084.

[26] O. Weller, K. Ricci, M. Marone, A. Chaffin, D. Lawrie, B. V. Durme, Seq vs seq: An open suite of paired encoders and decoders, 2025. URL: https://arxiv.org/abs/2507.11412. doi:10.48550/arXiv.2507.11412. arXiv:2507.11412, accepted to ICLR 2026.

[27] B. Warner, A. Chaffin, B. Clavié, O. Weller, O. Hallström, S. Taghadouini, A. Gallagher, R. Biswas, F. Ladhak, T. Aarsen, N. Cooper, G. Adams, J. Howard, I. Poli, Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and infer-

ence, 2024. URL: https://arxiv.org/abs/2412.13663. arXiv:2412.13663.

[28] JHU CLSP, jhu-clsp/ettin-encoder-400m, https://huggingface.co/jhu-clsp/ettin-encoder-400m, 2026. Hugging Face model card, accessed 2026-05-25.

[29] 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.