# Contrastive Windowed Boundary Classification for Multi-Author Writing Style Analysis

Notebook for the PAN Lab at CLEF 2026

Anastasia Voznyuk[1,*,†], Ahmed Tammaa[1,†]

[1]Technical University of Graz, Graz, Austria

**Abstract**

We describe the system submitted by team *DataTeam* to the PAN 2026 Multi-Author Writing Style Analysis shared task, which asks to predict authorship changes between consecutive pairs of sentences in a fanfiction document. We approach the task as windowed boundary classification: each boundary is encoded as a two-segment input with 2 sentences of left and right context, processed by a single DeBERTa-v3 encoder trained jointly with weighted binary cross-entropy and a supervised contrastive auxiliary loss. Short stylometric tag sentences are also added to each window as additional stylistic signal. We additionally tuned the data recipe and encoder warm-start separately for each difficulty, obtaining macro-F1 scores of **0.918**, **0.768** and **0.78** on the easy, medium and hard tracks; ranking first for medium and hard tracks of PAN 2026.

**Keywords**

Style change detection, Multi-author Analysis, Supervised contrastive learning, Boundary classification

## 1. Introduction

Authorship analysis at the sentence-pair level is the most fine-grained formulation of multi-author writing style detection. Instead of deciding whether a document has one or several authors, the model must locate every author transition. The PAN 2026 Multi-Author Writing Style Analysis (MAWSA) task makes this task even harder. First, the domain is fanfiction, where authors deliberately imitate shared characters and settings, so lexical and topical cues are weaker signals of authorship than in news or encyclopedic text. Second, the task is additionally divided into three difficulty levels that progressively reduce topic diversity between co-occurring authors, meaning the same surface topic can mask several distinct writing styles.

In this paper we describe the system submitted by the team *DataTeam*. The medium and hard submissions use a unified windowed-boundary architecture and vary only the data augmentation recipe and the encoder warm-start. Their shared components are: a context-expanded two-segment input over a DeBERTa-v3-base encoder; a small MLP classification head; an auxiliary projection head trained with supervised contrastive loss; stylometric tag sentences as a discrete side channel; and a two-stage post-processing pipeline (per-difficulty threshold tuning followed by a document-level authors calibrator). The easy submission reuses an earlier all-difficulty model with a different input format.

In the official PAN 2026 evaluation, our submission won the medium and hard tracks and finished 7th on easy.

## 2. Task Description

The MAWSA 2026 task asks to determine *authorship changes* between consecutive sentences in a multi-author document [1]. Similarly to the 2025 edition, this edition also increases the granularity of authors'

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

*Corresponding author.

†These authors contributed equally.

a.vozniuk@student.tugraz.at (A. Voznyuk); ahmed.tammaa@student.tugraz.at (A. Tammaa)

0009-0006-9679-6358 (A. Voznyuk); 0009-0005-9655-3770 (A. Tammaa)

input by shifting from paragraph-based authorship to sentence-based authorship. Organizers provided the dataset with three different levels of topic diversity to encourage participants to search for other signals of style change rather than simple topic information. The subtasks are described in a following way:

1. **Dataset 1 (Easy):** The sentences in a document cover various topics.
2. **Dataset 2 (Medium):** The sentences in a document cover fewer topics than the easy level.
3. **Dataset 3 (Hard):** All the sentences in a document cover the same topic.

The task is to identify all positions where the writing style changes within one text. Formally, for a document of $n$ sentences, the model must predict a binary vector $\mathbf{c} \in \{0, 1\}^{n-1}$, where $c_i = 1$ indicates that sentence $i+1$ was written by a different author than sentence $i$.

A new feature of this year's task edition is the updated domain of texts: organizers used an extensive collection of fanfiction texts, i.e., user-written stories that reuse characters and settings from existing works. This increases the difficulty of the task, as in fiction stories, author can try deliberately to change styles in dialogues or to convey something about particular character, and the overall topic diversity is not as pronounced as in scientific texts or news.

## 2.1. Dataset

The corpus contains 10,500 training and 2,250 validation documents per level (31,500 training documents in total). Table 1 summarises the key aggregate statistics.

**Table 1**
Dataset summary statistics (training split).

| Level | Docs | Avg sentences/doc | Avg authors/doc | Avg change rate |
|---|---|---|---|---|
| Easy | 10,500 | ≈53 | ≈3.0 | ≈0.305 |
| Medium | 10,500 | ≈124 | ≈3.0 | ≈0.044 |
| Hard | 10,500 | ≈57 | ≈3.0 | ≈0.209 |

Document length (measured in sentences) follows a right-skewed distribution across all three difficulty levels. The median document contains approximately 50–70 sentences, with a 25th–75th percentile range of roughly 30–100 sentences. A small number of documents exceed 130 sentences. The distributions are comparable across levels, confirming that difficulty derives from topic overlap rather than structural differences.

Each document is written by 2–5 authors, with the distribution approximately uniform across this range. No strong skew toward any particular author count was observed at any difficulty level, ensuring that the corpus tests the model across varying degrees of multi-authorship.

## 2.2. Label Imbalance

The most consequential finding for modeling is the class imbalance in the sentence-pair labels, particularly in the medium difficulty split. Table 2 reports the proportion of *same-author* (label 0) versus *changed-author* (label 1) pairs across all training sentence pairs.

**Table 2**
Label balance across difficulty levels (training split, all sentence pairs).

| Level | Total pairs | Same (0) | Changed (1) | % changed |
|---|---|---|---|---|
| Easy | 549804 | 382003 | 167801 | 30.52% |
| Medium | 1299324 | 1242693 | 56631 | 4.36% |
| Hard | 589946 | 466382 | 123564 | 20.94% |

The medium level exhibits the most pronounced imbalance, with only approximately 4% of pairs labeled as author changes. Then, we also noticed that median length of the same-author block is

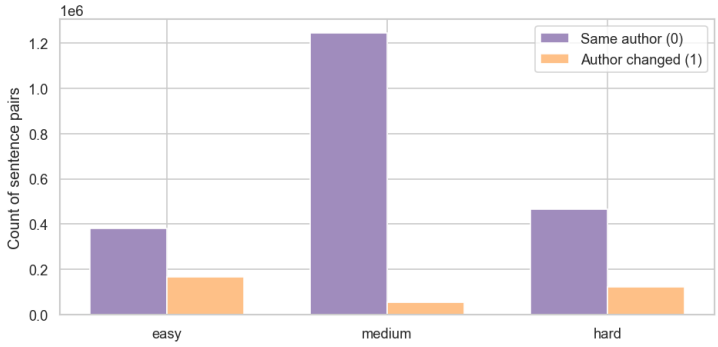


**Figure 1:** Label distribution in the training set.

only 1–4 sentences long across all difficulty levels. This has two practical implications: (i) authorship switches are locally detectable from a short context window, so encoding the full document at once is unnecessary; (ii) a model that over-predicts the "same author" will accumulate errors quickly because runs are short and changes are frequent.

## 3. Related Works

**Boundary classification for fine-grained authorship.** Localizing authorship transitions within a single document is a classical boundary-detection problem with close relatives in text segmentation, machine-generated text detection, and authorship verification. First, Lukasik et al. [2] framed topic-segmentation as a per-pair sentence-boundary classification problem with BERT [3], then Lo et al. [4] extended it to a hierarchical transformer over pretrained sentence encoders for neural text segmentation. Both works assume that the boundary signal comes from a content shift; and in that sense our task differs as the signal is purely stylometric, especially on medium and hard subtasks. Another common approach is to slide a window over the document, represent each window as a character $n$-gram profile, and detect anomalous segments via a style-change function [5], however these detectors are very sensitive to both style and content, and they are restricted for topic-controlled multi-author setup.

An adjacent direction of research is whether two text segments share the same author. One of such models, LUAR [6], is a RoBERTa model [7] trained with contrastive objectives on millions of Reddit posts, that can produce author-discriminative embeddings, transferable across domains. Then Wegmann et al. [8] disentangled style from topic in such representations, and Patel et al. [9] and Huertas-Tato et al. [10] further refined content-independent style embeddings via synthetic contrastive pairs. These representation-learning results motivate our use of a contrastively pretrained encoder to supply style-salient features to the boundary classifier.

In parallel, there are also works that tackle human-machine boundary detection: the SemEval-2024 Task 8 shared subtask [11] on machine-generated text boundary detection encourages a range of approaches centered on token-level sequence labeling, including DeBERTa-V3 [12] with data augmentation [13] or a decoder-encoder pipeline that first localizes and then refines the boundary [14]. Common to these systems is the use of a transformer encoder over local context, with CRF or calibration layers on top. Our windowed two-segment encoder plus authors calibrator follows the CRF idea, though framed at the sentence-pair and not on token level.

**Contrastive learning for writing-style change detection.** Within PAN itself, since PAN 2023 introduced topic-controlled splits, the strongest systems are usually some transformer encoders trained with an auxiliary contrastive objective. First, this idea was introduced by Chen et al. [15], that presented CoSENT-based style embeddings, and Ye et al. [16] with supervised contrastive learning over RoBERTa. The most direct predecessor of our work is SCL-DeBERTa [17], the PAN 2025 submission of Lin et al. that combines a DeBERTa backbone with a supervised contrastive auxiliary objective. Our system

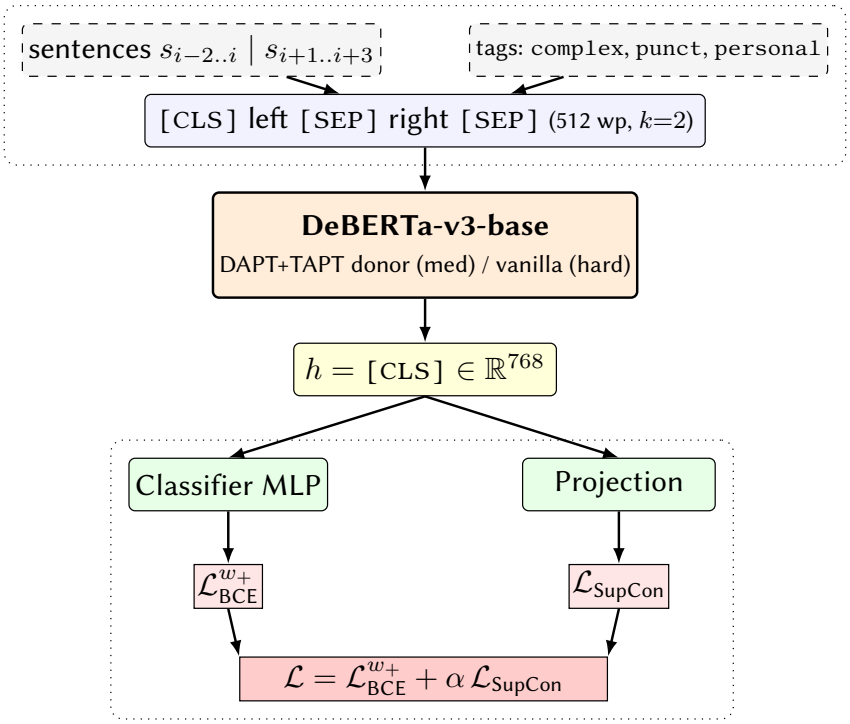


**Figure 2:** System architecture. Each boundary $(s_i, s_{i+1})$ is widened with $k$=2 context sentences and tokenized as a two-segment input; stylometric tags are added when the corresponding cue is present. A DeBERTa-v3-base encoder produces a [CLS] representation feeding two heads: an MLP classifier (weighted BCE) and a projection head (supervised contrastive loss).

inherits from SCL-DeBERTa the DeBERTa-v3 backbone, the supervised contrastive objective, and the auxiliary projection head, extending however this foundation with windowed two-segment inputs, stylometric tag side channels, and a DAPT+TAPT-adapted donor encoder for domain adaptation [18].

## 4. Proposed Method

The medium and hard submissions share a single architecture and differ mainly in the data recipe and encoder warm-start. The architecture (Figure 2) encodes each candidate boundary $(s_i, s_{i+1})$ as a two-segment input with $k = 2$ sentences of context on each side, passes it through a DeBERTa-v3-base encoder, and reads two heads off the [CLS] representation: an MLP classifier and a projection head used by the contrastive loss. Predictions are turned into final labels by a two-stage postprocessing pipeline. The medium and hard recipes (Sections 4.5.1–4.5.2) differ in whether the encoder is warm-started from a DAPT+TAPT donor and how the class imbalance is handled.

### 4.1. Input Representation

Each candidate boundary becomes a two-segment input: the left and right windows are tokenized with the standard DeBERTa-v3 tokenizer and concatenated as

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

with maximum of 512 tokens split between the two sides. We do not use any <BOUNDARY> special token in this configuration, as the two-segment encoding (`token_type_ids` together with the standard [SEP]) seemed like a sufficient signal for the boundary. When a window exhibits a surface stylistic cue, we prepend a short tag sentence as a discrete additional feature. Each of the three tags is computed and appended independently. Tags are emitted as keywords and placed at the start of each segment. The chosen tags are following:

- `sentence_complexity` defines whether text contains mostly short or long sentences, by measuring whether the average words-per-sentence is $\geq 21$ or $\leq 15$. If the length is between these thresholds, no complexity tag is added. The thresholds were set heuristically from the per-track average-length distribution.
- `punctuation` marks, whether text contains ellipsis, question marks, quotation marks, semicolons and when those marks occur. We ran an experiment where we checked the influence of punctuation on boundary changes and only these punctuation signs showed non-negligible signal.
- `personal_style` marks when the text contains first-person pronouns, capturing narrative point of view.

We chose these three families because they are classical stylometric markers that are largely topic-independent: sentence-length statistics, punctuation habits, and function-word usage remain discriminative even when topical cues are suppressed, as in the hard track. We also experimented with additional tags but observed no significant validation gain and did not retain them; selection was based on aggregate validation performance rather than per-tag ablation.

### 4.2. Model

A single DeBERTa-v3-base encoder produces $h \in \mathbb{R}^{768}$ from the `[CLS]` token. Two heads operate on $h$:

- **Classifier:** $\hat{y} = \sigma(W_2\,\mathrm{GELU}(W_1\,h))$, hidden width 256, dropout 0.1.
- **Projection (SCL only):** $g : \mathbb{R}^{768} \to \mathbb{R}^{128}$, Linear–GELU–Dropout–Linear.

Classification reads directly from $h$; the projection is used only by the contrastive loss.

### 4.3. Training Objective

The total loss combines a class-weighted binary cross-entropy with a supervised contrastive term:

$$\mathcal{L} = \mathcal{L}_{\mathrm{BCE}}^{w_+} + \alpha \cdot \mathcal{L}_{\mathrm{SupCon}} \tag{1}$$

$\mathcal{L}_{\mathrm{BCE}}^{w_+}$ is weighted BCE-with-logits with $w_+$ set from the empirical class prior and capped at a level-specific $w_+^{\max}$. It serves as a safety bound against the destabilizing effect of very large weights.

Following [19], we use supervised contrastive loss $\mathcal{L}_{\mathrm{SupCon}}$ to pull together pairs that share the same boundary label and push apart pairs that disagree, thus providing the encoder with a metric-style signal in addition to the per-pair BCE gradient. We apply an in-batch supervised contrastive loss over the projected, $\ell_2$-normalized `[CLS]` embeddings $z_i = g(h_i)/\|g(h_i)\|_2$, with temperature $\tau = 0.07$:

$$\mathcal{L}_{\mathrm{SupCon}} \;=\; -\frac{1}{|\mathcal{A}|}\sum_{i\in\mathcal{A}}\frac{1}{|P(i)|}\sum_{p\in P(i)}\log\frac{\exp(z_i\cdot z_p/\tau)}{\sum_{a\neq i}\exp(z_i\cdot z_a/\tau)}, \tag{2}$$

where $\mathcal{A}$ is the set of anchors in the batch and $P(i)$ is the set of in-batch peers sharing the boundary label of anchor $i$. The coefficient $\alpha$ for $\mathcal{L}_{\mathrm{SupCon}}$ was taken equal to 0.3. We also experimented with R-Drop, but it did not give us a noticeable increase in metric.

### 4.4. Postprocessing

(1) Per-difficulty thresholds are swept on the validation split over $[0.05, 0.95]$ in $0.01$ steps and the macro-F1 maximizer is kept. (2) An *authors calibrator* fits a linear $(\mathrm{scale}, \mathrm{offset})$ over a $17\times17$ grid mapping $\hat{N} = \sum_i p_{i,i+1}$ to the true author count, minimizing document-level MAE.

### 4.5. Level-specific Recipes

Table 3 summarizes the main differences between the medium and hard recipes; the rest is identical.

**Table 3**
Differences between the medium and hard recipes. Shared settings (input, model, loss form, $\tau$, $\alpha$, optimizer, postprocessing) are omitted.

| | **Medium** | **Hard** |
|---|---|---|
| Encoder init | DAPT+TAPT donor | vanilla DeBERTa-v3-base |
| Negatives | full ($\approx$ 22:1) | downsampled 3:1 |
| Cross-group aug. | off | on |
| $w_+^{\max}$ | 12.0 | 6.0 |

#### 4.5.1. Medium

We train on the full medium split at its natural $\approx$ 22:1 class ratio; imbalance is handled entirely at the loss level, which preserves the joint distribution the model will see at inference and keeps enough same-author pairs to support the SCL term. The encoder is warm-started from a DAPT+TAPT-adapted DeBERTa-v3-base trained on the PAN 2020 fiction corpus [20], run once as a separate stage and reused as a donor across runs. Pretraining on PAN 2020 fiction corpus exposed the encoder to general fiction-register statistics without letting it overfit to the specific 2026 corpus. At the natural ratio the automatic positive weight is above 20, which we found destabilizing, hence the $w_+^{\max} = 12.0$ cap.

#### 4.5.2. Hard

The hard recipe diverges from medium along the four axes of Table 3. First, *cross-group augmentation* synthesizes additional positive pairs by combining a sentence from block from the same author with the first sentence of the next authorship block. This augmentation is applied before downsampling so the synthetic positives count toward the positive pool. *Negative downsampling* caps negatives at three times the positive count, taking the 3.8:1 natural ratio down to $\sim$3:1 — mild enough that both classes appear in every batch without discarding much same-author signal. The vanilla encoder for hard is a deliberate choice to check, whether DAPT+TAPT is necessary. With the post-downsampling prior the auto $w_+ \approx 3$ is below 6.0.

#### 4.5.3. Easy

We did not run a level-specific recipe for easy and instead re-used a model trained on the union of difficulties. As a post-hoc check, we additionally trained an easy-track model using the same recipe as hard; this raised macro-F1 only marginally, to 0.922. The small gain probably indicates that the easy track is already saturated, so the windowed-context and contrastive frameworks adds little gain. We therefore report the reused model as our primary easy result.

## 5. Results and Analysis

Systems were evaluated by the macro-averaged F1 score over the binary per-pair labels on the TIRA platform [21]. For each document, predictions $\hat{\mathbf{c}}$ are compared to the gold change vector $\mathbf{c} \in \{0, 1\}^{n-1}$, and the F1 of the same-author (0) and change (1) classes is averaged; the final score is the mean over all documents.

Table 4 reports our official scores against the shared-task baseline. Our system reached macro-F1 of 0.918, 0.768, and 0.78 on easy, medium, and hard, ranking 7th, 1st, and 1st.

The "hard" and "medium" tracks conflate two orthogonal axes. The intended axis is topic diversity, with available topic signal decreasing from easy to hard. The second, incidental axis is label imbalance: medium is by far the most skewed, with only 4.36% positive pairs versus 20.94% (hard) and 30.52% (easy) (Table 2). Under macro-F1, extreme imbalance depresses the minority-class F1 and thus the average. Medium is therefore bottlenecked by imbalance while hard is bottlenecked by the absence of topic cues, and the near-tie in scores between them is the result of these two distinct effects.

Easy and hard have comparable positive rates (30.5% vs. 20.9% in Table 2), so differences in imbalance between them are much smaller. The ≈0.14 macro-F1 gap (0.918 vs. 0.78) is thus largely attributable to the removal of topic cues. This both confirms that the model exploits topic information when it is available and quantifies how much performance depends on it. It is also consistent with the marginal +0.004 gain of the full windowed-contrastive recipe on easy (stone-valley, Table 4): when topic cues are already strong, the added stylistic tags do not contribute a lot.

**Table 4**
Official PAN 2026 results for the submitted DataTeam system. The system ranked 1st on the medium and hard tracks; the post-hoc run (stone-valley) is shown for comparison.

| Team | Submission name | Task1 F1 | Task2 F1 | Task3 F1 |
|---|---|---|---|---|
| datateam-tug | (brown-causeway) | 0.918 | 0.768 | 0.78 |
| datateam-tug | (stone-valley) | 0.922 | 0.768 | 0.78 |
| baseline | (brass-bagging) | 0.41 | 0.489 | 0.441 |

## 6. Conclusion

We presented *DataTeam*'s PAN 2026 system. The medium and hard submissions use a context-expanded supervised contrastive DeBERTa framework that builds on the SCL-DeBERTa recipe [17]. Our novel changes include widening each boundary into a two-segment $k = 2$ window, appending stylometric tags of the sentences as a discrete side channel, and using a DAPT+TAPT-adapted donor encoder. The system achieved macro-F1 of **0.768** and **0.78** on the medium and hard tracks of PAN 2026, both ranking first.

## 7. Limitations

Our experimental design does not allow to fully break down each individual component's contribution. The medium and hard recipes differ simultaneously in encoder warm-start, negative sampling, cross-group augmentation, and the positive-weight cap (Table 3), so the contribution of any single factor cannot be isolated from the others. We report no controlled ablations due to technical difficulties in testing models after the competition.

The stylometric tags rely on heuristic thresholds and a hand-selected punctuation subset; they are surface-level and English- and register-specific. We experimented with other stylistic tags, but we did not observe a significant gain on validation, even though they might be useful for other registers and languages. We also did not quantify the individual effect of any stylistic tags. So this limits the conclusions we can draw about generalization to other domains, languages, and longer-range stylistic dependencies.

## Declaration on Generative AI

During the preparation of this work, the author(s) used Claude and Gemini for grammar check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

## References

[1] J. Bevendorff, M. Fröbe, A. Greiner-Petter, A. Jakoby, M. Mayerl, P. Nakov, H. Plutz, M. Potthast, B. Stein, M. N. Ta, Y. Wang, E. Zangerle, in: M. Hagen, M. Potthast, B. Stein, P. Schaer, E. Zangerle,

A. Barrón-Cedeño, A. G. S. de Herrera, E. S. Salido, S. MacAvaney, J. M. Struß (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2026.

[2] M. Lukasik, B. Dadachev, K. Papineni, G. Simões, Text segmentation by cross segment attention, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 4707–4716. URL: https://aclanthology.org/2020.emnlp-main.380/. doi:10.18653/v1/2020.emnlp-main.380.

[3] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), 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, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423/. doi:10.18653/v1/N19-1423.

[4] K. Lo, Y. Jin, W. Tan, M. Liu, L. Du, W. Buntine, Transformer over pre-trained transformer for neural text segmentation with enhanced topic coherence, in: M.-F. Moens, X. Huang, L. Specia, S. W.-t. Yih (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 3334–3340. URL: https://aclanthology.org/2021.findings-emnlp.283/. doi:10.18653/v1/2021.findings-emnlp.283.

[5] E. Stamatatos, Intrinsic plagiarism detection using character n-gram profiles, in: Proceedings of the 3rd International Workshop on Uncovering Plagiarism, Authorship, and Ownership (PAN), 2009, pp. 41–46.

[6] R. A. R. Soto, O. Miano, J. Ordonez, B. Chen, A. Khan, M. Bishop, N. Andrews, Learning universal authorship representations, in: EMNLP, 2021.

[7] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).

[8] 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, 2022, pp. 249–268.

[9] A. Patel, J. Zhu, J. Qiu, Z. Horvitz, M. Apidianaki, K. McKeown, C. Callison-Burch, Styledistance: Stronger content-independent style embeddings with synthetic parallel examples, in: Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), 2025, pp. 8662–8685.

[10] J. Huertas-Tato, A. Martin, D. Camacho, Part: Pre-trained authorship representation transformer, 2025. URL: https://arxiv.org/abs/2209.15373. arXiv:2209.15373.

[11] Y. Wang, J. Mansurov, P. Ivanov, J. Su, A. Shelmanov, A. Tsvigun, O. Mohammed Afzal, T. Mahmoud, G. Puccetti, T. Arnold, SemEval-2024 task 8: Multidomain, multimodel and multilingual machine-generated text detection, in: A. K. Ojha, A. S. Doğruöz, H. Tayyar Madabushi, G. Da San Martino, S. Rosenthal, A. Rosá (Eds.), Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 2057–2079. URL: https://aclanthology.org/2024.semeval-1.279/. doi:10.18653/v1/2024.semeval-1.279.

[12] P. He, J. Gao, W. Chen, Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing, 2021. arXiv:2111.09543.

[13] A. Voznyuk, V. Konovalov, DeepPavlov at SemEval-2024 task 8: Leveraging transfer learning for detecting boundaries of machine-generated texts, in: A. K. Ojha, A. S. Doğruöz, H. Tayyar Madabushi, G. Da San Martino, S. Rosenthal, A. Rosá (Eds.), Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 1821–1829. URL: https://aclanthology.org/2024.semeval-1.257/.

doi:10.18653/v1/2024.semeval-1.257.
[14] A. Shirnin, N. Andreev, V. Mikhailov, E. Artemova, AIpom at SemEval-2024 task 8: Detecting AI-produced outputs in m4, in: A. K. Ojha, A. S. Doğruöz, H. Tayyar Madabushi, G. Da San Martino, S. Rosenthal, A. Rosá (Eds.), Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 1667–1672. URL: https://aclanthology.org/2024.semeval-1.238/. doi:10.18653/v1/2024.semeval-1.238.
[15] H. Chen, Z. Han, Z. Li, Y. Han, A Writing Style Embedding Based on Contrastive Learning for Multi-Author Writing Style Analysis, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2023, pp. 2562–2567. URL: https://ceur-ws.org/Vol-3497/paper-206.pdf.
[16] Z. Ye, C. Zhong, H. Qi, Y. Han, Supervised Contrastive Learning for Multi-Author Writing Style Analysis, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2023, pp. 2817–2822. URL: https://ceur-ws.org/Vol-3497/paper-237.pdf.
[17] K. Lin, C. Liu, F. Ye, Z. Han, SCL-DeBERTa: Multi-Author Writing Style Change Detection Enhanced by Supervised Contrastive Learning, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, volume 4038, CEUR-WS.org, 2025, pp. 3781–3787. URL: https://ceur-ws.org/Vol-4038/paper_302.pdf.
[18] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, N. A. Smith, Don't stop pretraining: Adapt language models to domains and tasks, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 8342–8360. URL: https://aclanthology.org/2020.acl-main.740/. doi:10.18653/v1/2020.acl-main.740.
[19] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, D. Krishnan, Supervised contrastive learning, Advances in neural information processing systems 33 (2020) 18661–18673.
[20] J. Bevendorff, B. Ghanem, A. Giachanou, M. Kestemont, E. Manjavacas, I. Markov, M. Mayerl, M. Potthast, F. Rangel, P. Rosso, G. Specht, E. Stamatatos, B. Stein, M. Wiegmann, E. Zangerle, Overview of PAN 2020: Authorship Verification, Celebrity Profiling, Profiling Fake News Spreaders on Twitter, and Style Change Detection, in: A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, H. Joho, C. Lioma, C. Eickhoff, A. Névéol, L. Cappellato, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. 11th International Conference of the CLEF Initiative (CLEF 2020), volume 12260 of *Lecture Notes in Computer Science*, Springer, Berlin Heidelberg New York, 2020, pp. 372–383. doi:10.1007/978-3-030-58219-7_25.
[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.

## A. Hyperparameters

Shared training settings across both runs are given in Table 5. Decoupled learning rates reflect the warm-started encoder in the medium recipe (gentle updates over an already-adapted backbone vs. larger steps for the fresh head and projection); the same schedule is kept for hard to keep the two runs comparable.

**Table 5**
Shared training hyperparameters.

| | |
|---|---|
| Precision | bf16 |
| Micro-batch / grad. accum. | 16 / 1 |
| Encoder LR | $1 \times 10^{-5}$ |
| Head + projection LR | $1 \times 10^{-4}$ |
| Warmup | 2% of training steps |
| Weight decay | 0.01 |
| Max gradient norm | 1.0 |
| Epochs | up to 10 |
| Early stopping | patience 3 on PAN-tuned macro-F1 |