Team avocadobeans at PAN 2026: Style Change Detection with DeBERTa, Evidence Aggregation and Sentence Length Embedding

Notebook for the PAN Lab at CLEF 2026

Łukasz Gągała[1]

[1]Universität Göttingen Seminar für Deutsche Philologie Jacob-Grimm-Haus Käte-Hamburger-Weg 3 37073 Göttingen, Germany

Abstract

Detection of style change in multi-author texts is a challenge not only in the domains of computational linguistics and forensics but also for applications like intelligent writing assistants. In this paper we present our proposal for 2026 edition of PAN Style Change Shared Task that we developed from the successful ideas of the last year's edition. The task consists of three levels of difficulty (`easy`, `medium`, `hard`), for which there are separate training and validation subsets of data. We describe enhancements brought in and additional data we use to train the final model and the final results thereof.

Keywords

authorship verification, computational linguistics, writing style change detection, DeBERTa, PAN 2026

1. Introduction

Authorship verification is one of the main concerns of forensic linguistics. A particular case of this problem is writing style change detection, in which a single document may have been authored by two or more individuals, and the goal is to identify the boundaries between different authors' contributions. Research on this task has been actively promoted through a PAN Shared Task at the CLEF conferences for many years [1]. In the 2026 edition [2] participants are asked, as in previous editions, to develop an approach for detecting a potential authorship change between two consecutive sentences across multiple documents. The dataset provided to participants consists of training and validation subsets, while the test dataset is evaluated within the TIRA environment [3]. Three levels of difficulty are considered: easy, medium, and hard. These levels reflect varying degrees of topic consistency, which may facilitate the detection of writing style changes.

2. Related Work

Two most successful approaches of the 2025 edition of PAN Style Change Shared Task were based on DeBERTa architecture [4]. [5] used it as an encoder for a two layer classifier with `CLS` token. On the other hand [6] deployed two pathways with the classification head learning a joint objective with supervised contrastive loss – besides the style change probability it learnt also a style vector. Other teams of the 2024 PAN edition also explored capabilities of different models, e.g. [7] evaluated ELECTRA, SqueezeBERT, and RoBERTa in a systematic way and [8] incorporated RoBERTa, DeBERTa, and ERNIE models into a single ensemble approach. Outside the context of the PAN Shared Task, authorship verification has also attracted growing research interest. [9] built a large encoder model for author style by fine-tuning RoBERTa on multilingual dataset consisting of millions of text samples from Wikipedia, Reddit, Public Domain books and other sources. The model uses a contrastive loss to embed thousands of authors in one representation space. We make use of samples from a text corpus

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

lukaszgagala@wp.pl (Ł. Gągała)

https://github.com/lukasz-g (Ł. Gągała)

Table 1
Distribution of training pairs derived from the PAN 2026 Style Change dataset

Subset	Label (change=1|no-change=0)	Number of sentences
easy	no-change	459339
easy	change	268678
medium	no-change	1385260
medium	change	61958
hard	no-change	593102
hard	change	134237

Table 2
Distribution of additional training pairs derived from Wikipedia and fan-fiction datasets

Subset	Label (change=1|no-change=0)	Number of sentences
fanfic	no-change	50000
fanfic	change	50000
wiki	no-change	50000
fanfic	change	50000

used for the training of that model. [10] compiled a corpus of fan fiction texts in over 40 languages to train a model for authorship verification. In our approach we turn to English language part of their training corpus.

3. Data

The PAN Shared Task dataset consists of 3 subsets with varying level of difficulty: `easy`, `medium` and `hard`. Each subset is a collection of text documents; each document contains a various number of sentences written by at least one author. The task is to identify the boundary between each author if present. The sentences in the `easy` subset cover a range of topics, allowing the participants to key off topic information when detecting authorship changes. In the `medium` subset the relatively small amount of topical variation leaves approaches with fewer topical cues to draw on, causing them to rely more heavily on style. All sentences in the `hard` subset are on the same topic, so that any method must based solely on stylometric features. From all the subsets we constructed training pairs. Their distribution is presented in Table 1.

As the dataset for this year's edition of Shared Task is significantly larger than the previous one, we decided also to collect and use training samples from two large corpora for authorship verification:

- **Wikipedia authors.** We selected 100k samples of training passage pairs for English language covering both Wikipedia articles and discussions among Wikipedians [9] [1]. The class balance was maintained.
- **Fanfiction authors.** From the fan fiction corpus we drew 100k samples in English with the balance for `change` and `no-change` classes.

4. Architecture

The architecture of our style change classifier consists of two main components: an encoder and a classification head. We enhance the standard pipeline by unfreezing the last two layers of the encoder, adding a gated fusion layer, and incorporating additional embeddings for text length and domain. An overview of the architecture is presented in Figure 1.

[1]https://huggingface.co/datasets/Blablablab/MAC

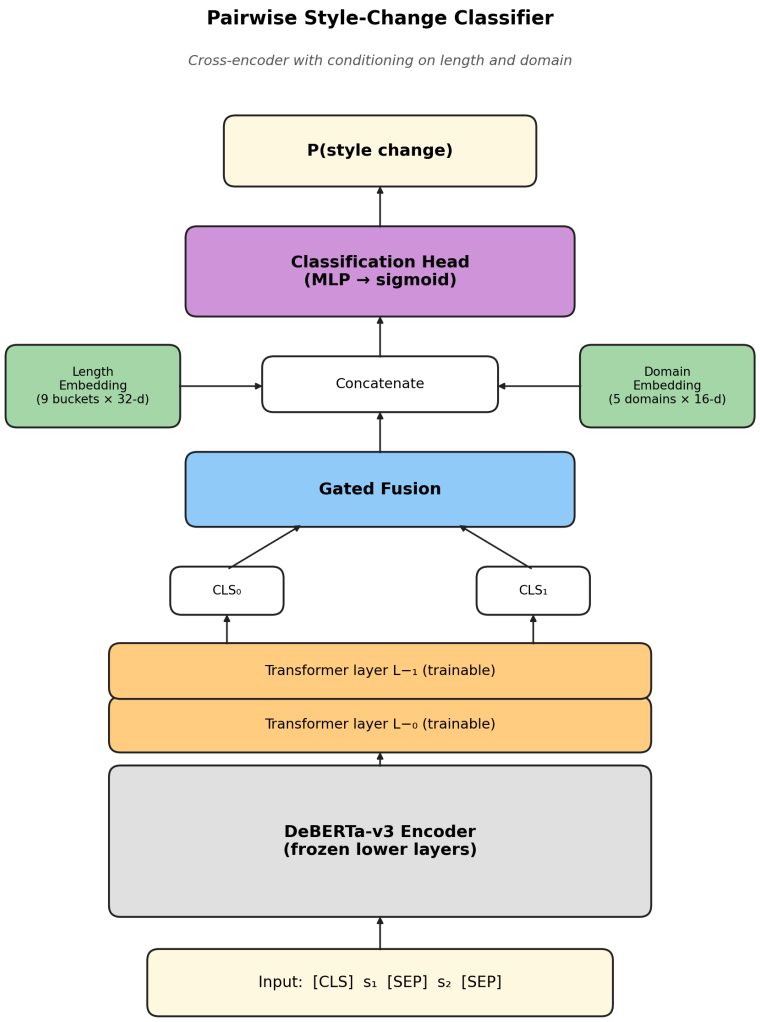

Figure 1: Overview of architecture with DeBERTa as encoder.

4.1. Encoder

As it was proven by [5] and [6] the DeBERTa model [11] can be a successful encoder for author style change classifiers, primarily due to its capability of disentanglement of positional encoding what is believed to better catch nuanced authorial footprint. The markers of style change do not need to be strictly related to word positions and DeBERTa seems to capture that phenomenon. In our work we adopted `deberta-v3-large` version of the model [2].

4.2. Unfrozen Layers

To further build on the capabilities of the transformer architecture, we unfreeze the final two layers of the encoder (out of a total of 24 layers) and fine-tune them jointly with the classification head (see below). This allows the encoder representations to better adapt to the downstream classification task by steering the learned parameters toward a style-change-specific representation space. We take the outputs from the final two layers of the DeBERTa encoder and fuse them through a Gated Fusion Layer [12]. Before passing the resulting fused sentence-pair representation on to the final classification head, we augment it with two additional embeddings that capture information about text length and domain

[2]https://huggingface.co/microsoft/deberta-v3-large

Table 3
Hyperparameter settings of our model

Hyperparameter	Value
max length	509
dimensions of length embedding	32
dimensions of domain embedding	32
head dimensions	1024+32+32
dropout	0.1
batch size	128
initial learning rate	1×10^{-5}
epochs	5
weight decay	0.001
number of unfrozen layers	2

category.

4.3. Embeddings for Text Length and Data Source

As we conducted analysis of sentence samples both in PAN dataset and additional datasets of Wikipedia and fan-fiction, we discovered that those additional training data represent a different distribution of length and mixing them together caused harm to model performance for PAN subsets. As the input length can be critical for transformers and different mechanisms can be used to classify e.g. a sentence of 16 tokens vs a text passage of 256 tokens. To facilitate length sentence recognition, we decided to add auxiliary embedding for each input sentence pair. The model sets up its classification head around input length so it can pick up on authorship signals and mollify differences between short and long passages. We bucketed each sentence's token count into one of nine discrete bins whose boundaries (16, 24, 32, 48, 64, 96, 128, 192, 256) are spaced log-linearly to allocate finer resolution where short-text signal degradation is most pronounced, and map each bin to a learnable 32-dimensional embedding. To let the model adapt to systematic differences across our heterogeneous training mix, we attach learnable embeddings of 32 dimensions for each data domain (5 categories: `PAN-easy`, `PAN-medium`, `PAN-hard`, `fanfic`, `wiki`) that captures shared statistical characteristics. The final sentence-pair representation concatenates these length and domain embeddings with the fused encoder output before the classification layer, allowing the head (two dense layers with dropout and ReLu activation) to factor in the available stylistic evidence when drawing up its decision boundary.

4.4. Classification Head

The classification head for binary classification was slightly enhanced in comparison with [5]and [6]. We added a fusion layer that was supposed to better compress information from more than one encoder layers.

4.5. Settings

Table 3 summarises the training configuration used in our experiments. As our model introduces several additional components compared with [5]and [6], we intentionally stick to the default hyper-parameter settings wherever possible, leaving their optimisation for future work to build on.

5. Test Time Sample Merging

As the sentence samples in the PAN datasets do not fully utilise the input length supported by the DeBERTa model, we experimentally investigated combining multiple sentences into a single input during inference (see Figure 2). The underlying idea was to perform the classification iteratively over each

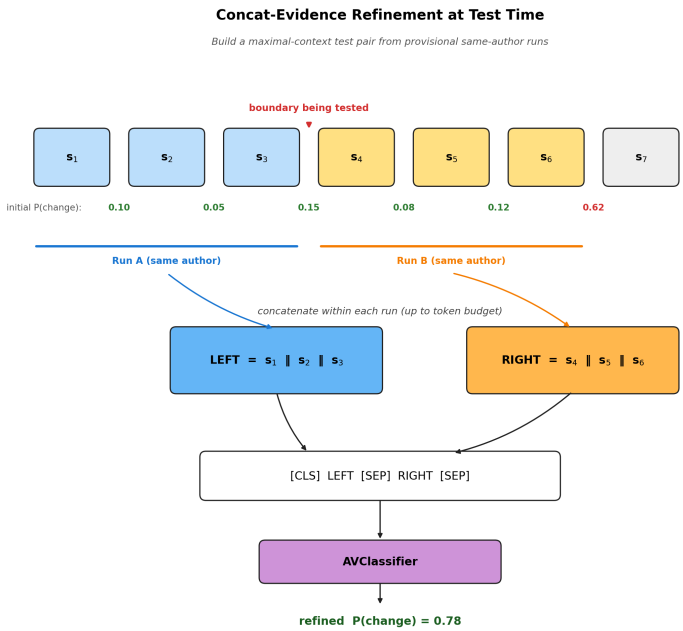

Figure 2: Merging Consecutive Sentences for Evidence Aggregation.

Table 4
Results on the test dataset obtained in the TIRA environment.Performance is reported as the $F_{1,\text{macro}}$ score. Our results outperform both baselines but remain more than 0.1 points behind the best-performing teams.

Team	Easy F_1	Medium F_1	Hard F_1
stylefusion	1.000	0.767	0.775
Almoment	1.000	0.741	0.773
ours	0.918	0.617	0.664
Baseline Predict 1	0.234	0.042	0.173
Baseline Predict 0	0.410	0.489	0.441

document for N passes, merging sentences that had been classified as written by the same author in the preceding pass. We hypothesised that increasing the value of N would improve classification performance on the validation data. However, our preliminary experiments did not yield substantial improvements, and we therefore fixed the number of concatenation passes to one. Further investigation of sentence merging is left for future work.

6. Results

Table 4 presents the results obtained in the TIRA environment. Our system trails the best-performing team by 0.1 F_1 points across all subset categories. Given the scope of our work, we believe there is considerable potential for further improving the model. Notably, the results on the easy subset indicate that an F_1 score of 1.0 is achievable, highlighting the potential for further research and optimisation.

7. Future Work

In future work, we intend to investigate more extensively whether (a) adding and/or unfreezing additional encoder layers can bring about improvements in model performance; (b) the conditions under

which merging sentences as evidence of style change results in better performance; and (c) whether incorporating additional out-of-domain training samples can give rise to further performance gains.

Declaration on Generative AI

During the preparation of this work, the author used GPT-5.5 in order to paraphrase and reword, to improve writing style and to check grammar and spelling. Further, the author used Claud Opus-4.8 for Figures 1 and 2 in order to generate images. After using these tools, the author reviewed and edited the content as needed and takes 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] 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.

[3] 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.

[4] E. Zangerle, M. Mayerl, M. Potthast, B. Stein, Overview of the multi-author writing style analysis task at pan 2025, in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025), Madrid, Spain, 9-12 September, 2025, volume 4038 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2025, pp. 3568–3574. URL: https://ceur-ws.org/Vol-4038.

[5] X. Lin, C. Liu, X. Duan, Z. Han*, Team wqd at style change detection in multi-author writing: A deep learning approach based on deberta, in: Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, 2025. URL: https://ceur-ws.org/Vol-4038/paper_303.pdf.

[6] K. Lin, C. Liu, F. Ye, Z. Han, Scl-deberta: Multi-author writing style change detection enhanced by supervised contrastive learning, in: Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, 2025. URL: https://ceur-ws.org/Vol-4038/paper_302.pdf.

[7] T. Lin, Y. Wu, L. Lee, Team nycu-nlp at pan 2024: Integrating transformers with similarity adjustments for multi-author writing style analysis, CEUR Workshop Proceedings 3740 (2024) 2716–2721. Publisher Copyright: © 2024 Copyright for this paper by its authors.; 25th Working Notes of the Conference and Labs of the Evaluation Forum, CLEF 2024 ; Conference date: 09-09-2024 Through 12-09-2024.

[8] 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: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF), volume 3740 of *CEUR Workshop Proceedings*, 2024, pp. 2688–2694.

[9] A. Israeli, S. Liu, J. May, D. Jurgens, The million authors corpus: A cross-lingual and cross-domain Wikipedia dataset for authorship verification, in: W. Che, J. Nabende, E. Shutova, M. T. Pilehvar (Eds.), Findings of the Association for Computational Linguistics: ACL 2025, Association for Computational Linguistics, Vienna, Austria, 2025, pp. 25997–26017. URL: https://aclanthology.org/2025.findings-acl.1335/. doi:10.18653/v1/2025.findings-acl.1335.

[10] S. Bischoff, N. Deckers, M. Schliebs, B. Thies, M. Hagen, E. Stamatatos, B. Stein, M. Potthast, The

importance of suppressing domain style in authorship analysis, 2020. URL: https://arxiv.org/abs/2005.14714. arXiv:2005.14714.

[11] P. He, X. Liu, J. Gao, W. Chen, Deberta: Decoding-enhanced BERT with disentangled attention, CoRR abs/2006.03654 (2020). URL: https://arxiv.org/abs/2006.03654. arXiv:2006.03654.

[12] J. Arevalo, T. Solorio, M. Montes-y Gómez, F. A. González, Gated multimodal networks, Neural Comput. Appl. 32 (2020) 10209–10228. URL: https://doi.org/10.1007/s00521-019-04559-1. doi:10.1007/s00521-019-04559-1.