Overview of the Multi-Author Writing Style Analysis Task at PAN 2025

Eva Zangerle[1], Maximilian Mayerl[2], Martin Potthast[3] and Benno Stein[4]

[1]*University of Innsbruck*
[2]*University of Applied Sciences BFI Vienna*
[3]*University of Kassel, hessian.AI, and ScaDS.AI*
[4]*Bauhaus-Universität Weimar*

pan@webis.de   https://pan.webis.de

Abstract

The multi-author writing style analysis task at PAN 2025 aims at identifying the exact positions of writing style changes within documents written by multiple authors. This is a crucial step for further tasks such as authorship attribution, plagiarism detection, or the identification of gift authorships. In the 2025 edition, we ask participants to detect style changes at the sentence level across three subtasks, thereby advancing the task towards realistic, real-world scenarios. The datasets differ in topical homogeneity, but also in sentence-level similarity, which allows controlling the difficulty of the task. This paper presents an overview the task, describes the dataset provided, summarizes the approaches submitted by participants, and discusses the results obtained.

1. Introduction

The task of analyzing multi-author documents consists of computing stylistic profiles of the authors on text features that capture their characteristic styles. These profiles are computed solely through intrinsic analysis, without resorting to external sources or reference corpora. By comparing these profiles, changes in style—and possibly also in authorship—can be detected. This task is therefore fundamental for several downstream applications such as authorship attribution or text reuse detection.

The multi-author writing style analysis task has been part of PAN since 2016. Since then, the tasks and datasets have been refined, while observing a paradigm shift in the approaches submitted by the participants. In the first years, the task was to cluster text segments by author [1]. In subsequent years, tasks were distinguished to determine whether a given document was written by one or more authors [2, 3] and to indicate the positions of style changes [3]. In 2019, participants were asked not only to perform this binary classification, but also to predict the number of authors of the given document [4]. As of the 2020 edition, participants had to detect style changes at the paragraph level for the first time (i.e., detect whether there is a style change between two consecutive paragraphs) [5, 6]. In 2022, the sentence level was added [7], whereas in 2023 and 2024, only the paragraph level [8, 9] was considered, albeit for three datasets with increasing difficulty.

In recent years, we have seen a dramatic paradigm shift in the field of writing style analysis. In early versions of our task, the most important steps consisted of extracting lexical, syntactic, and structural features [10] from parts of a document, which were then fed into a classifier to determine whether or not a style change had occurred. However, we have observed a shift towards large language models that are fine-tuned on the (augmented) training set [11, 12, 13, 14, 15, 16].

For the 2025 edition of the multi-author writing style analysis task, we ask participants to identify writing style positions at the sentence level. For each pair of consecutive sentences, the task is to determine whether the writing style changes between them. This is a binary classification problem. In previous editions, changes in writing style were handled with a coarser granularity, mostly focusing on paragraphs. This allowed models to compare larger text contexts, making it easier to identify stylistic differences based on more information. The renewed transition to recognizing style changes at the

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

sentence level increases the difficulty of the task: the style must now be derived from a smaller context window—a pair of two sentences. While this makes the task more difficult, it also brings it closer to real-world applications, where style changes can occur anywhere in the document.

In the Section 2, the task, the datasets provided to the participants, and the evaluation setup are described in detail. Section 3 summarizes the approaches submitted by the participants, and Section 4 presents and discusses the results of the twelve participating teams. Section 5 concludes the paper.

2. Style Change Detection Task

This section outlines the task, the datasets used, and the evaluation setup for the task.

2.1. Task Definition

The multi-author writing style analysis task as part of PAN 2025 requires participants to identify style changes within a given text. Specifically, the task is to determine whether there is a change in writing style between each pair of consecutive sentences. Once again, we are aiming for a more realistic scenario, reducing the detection of style changes at the paragraph level in the past year to the sentence level. Participants received three datasets that differ in terms of the difficulty of detecting style changes:

- **Easy.** Each document covers multiple topics, enabling participants to use topic shifts as a strong cue for detecting style changes. Additionally, the sentences within each document exhibit relatively low stylistic similarity.
- **Medium.** Documents are more topically coherent, requiring participants to rely more on stylistic features than on topic shifts. The stylistic similarity between sentences is moderate.
- **Hard.** All sentences in a document pertain to a single topic and are stylistically similar, removing topic variation as a helpful signal.

2.2. Dataset

The dataset for the 2025 edition of the task is once again based on posts by users on Reddit, a popular social media platform where all kinds of topics can be discussed. The discussions take place in "subreddits,' which are individual communities that focus on specific topics. For our datasets, we selected four subreddits where the discussions are particularly comprehensive due to their topics: *r/worldnews*, *r/politics*, *r/askhistorians*, and *r/legaladvice*.

We collected the individual threads from the subreddits mentioned and performed preprocessing steps such as removing quotes, markdown, emojis, hyperlinks, multiple line breaks, and extra spaces. Each post within the collected threads is then split into sentences. For each sentence, we calculate stylistic and semantic feature vectors. This allows us to merge individual sentences into the final documents in the next step. In particular, the feature vectors enable the calculation of both the topic (semantic) and stylistic similarity between individual sentences. This allows us to control the difficulty of the task by applying different similarity thresholds for pairs of consecutive sentences. We configure the similarity threshold for consecutive sentences so that it is (1) relatively high for the easy dataset, (2) moderate for the medium dataset, and (3) low for the difficult dataset.

Each of the three datasets comprises 6,000 documents. For all of them, we provided participants with training, validation, and test splits: 70% of the documents were used for training, 15% each for validation and testing. The test sets were held back and only used in the evaluation phase of the competition.

2.3. Evaluation

Participants are required to submit their code to the TIRA platform for evaluation and optimization. The submitted approaches are evaluated using the held back test data using the F_1-Measure measure. We calculate the F_1-Measure for each dataset individually and determine the macro-averaged score across all documents in the data set.

3. Survey of Submissions

We received twelve submissions for this year's edition of the task, with each participant submitting both their software and a notebook paper. Below is a brief description of each submission.

Alsheddi and Menai [17] propose using a graph convolutional network to solve this task. They represent each document as a graph, where the nodes represent the sentences in the document and the edges represent the boundaries between adjacent sentences. The goal of this approach is to model the boundaries—and thus also style changes—between sentences more explicitly. Text features are extracted using the pre-trained STAR model and used as the initial representation for the graph nodes. Edge representations are then learned using the graph convolutional network.

Boriceanu and Băltoiu [18] use an approach based on word adjacency networks to solve the task. Sentences in each document are represented as graphs, where nodes represent words in the sentence and edges represent consecutive words. Additional nodes and edges are added to represent the grammatical structure of the sentence at the part-of-speech level. Graph features are then extracted from these word adjacency networks and form a feature set that is used for the actual classification of style changes, combined with other features such as sentence-level embeddings and classic stylometric features. The actual detection of style changes is then performed by a gradient boosting classifier that uses these features as input.

Meier et al. [19] use an ensemble of models. They optimize a language model separately for each difficulty level and then combine these models to obtain predictions, with the combination being based on weights that are automatically derived for each sentence pair. This allows the model to operate without knowledge of the difficulty level of individual cases, making their approach more applicable to real-world scenarios where this information is not available.

Hosseinbeigi and Mehrani [20] also chose an approach based on an ensemble of two models, with the stated goal of having the two models capture different characteristics of the text. Their first model is based on a fine-tuned LaBSE model and is primarily designed to capture semantic information from the text. Their second model consists of a Siamese B-LSTM and is tasked with capturing mainly morphological and structural aspects. The results of these two models are then passed on to an XGBoost classifier to obtain the final results.

Chen et al. [21] use the pre-trained Llama-3 model to solve the task. To make both the model itself and its fine-tuning less computationally intensive, they apply 4-bit quantization to the model and use the IA3 fine-tuning method for more efficient tuning. Their approach thus focuses more on efficiency than on predictive performance.

Lin et al. [16] use the DeBERTa model as the basis for their approach. To fine-tune this model for the task, they use Bayesian optimization to search for optimal training hyperparameters.

Schmidt et al. [22] propose using a model based on the Bi-LSTM architecture. The goal is to explicitly model the sequential nature of documents—and thus also the writing style—so that the broader context in which style changes occur can be better taken into account. Initial embedding representations for each sentence were obtained using a fine-tuned StyleDistance model.

Liu et al. [23] propose contrastive learning to solve the task. It uses Llama-3 as a base model, combined with low-rank adaptation to make fine-tuning computationally more feasible. In addition, BERT-MLM is used to generate adversarial samples and expand the training data.

Księżniak et al. [24] focus on punctuation as stylistic markers to solve the task. They use contrastive learning to train a series of encoder models, each focusing on a specific punctuation pattern. These models were then fine-tuned to the actual task of detecting style changes at sentence boundaries.

Bölöni-Turgut et al. [25] use an ensemble of several language models to solve the task. The individual models are fine-tuned and combined with a feedforward neural network to obtain classifiers for style changes. The classifiers are then combined into an ensemble.

Lin et al. [26] use a supervised contrastive learning approach to obtain a model that can better distinguish between writing styles, using DeBERTa as the underlying language model.

Rohra et al. [27] use a fine-tuned RoBERTa model to solve the task.

Table 1
Overall results for the multi-author writing style analysis task, sorted by average F_1 across all three datasets. The best results are marked in bold.

Team	Team name	Easy F_1	Medium F_1	Hard F_1
Lin et al. [16]	wqd	0.958	0.823	**0.830**
Lin et al. [26]	xxsu-team	0.955	**0.825**	0.829
Boriceanu and Băltoiu [18]	stylospies	**0.959**	0.786	0.791
Hosseinbeigi and Mehrani [20]	team-tmu	0.950	0.792	0.792
Schmidt et al. [22]	better-call-claude	0.929	0.815	0.731
Księżniak et al. [24]	openfact	0.919	0.771	0.752
Bölöni-Turgut et al. [25]	cornell-1	0.909	0.793	0.698
Rohra et al. [27]	batatavada-pict	0.823	0.766	0.667
Meier et al. [19]	hhu	0.761	0.666	0.642
Alsheddi and Menai [17]	ksu	0.507	0.747	0.467
Chen et al. [21]	hellojie	0.461	0.583	0.484
Liu et al. [23]	team-of-bf	0.486	0.443	0.473
Baseline Predict 1		0.178	0.177	0.147
Baseline Predict 0		0.439	0.440	0.453

4. Evaluation Results

Table 1 shows the F_1 scores for the three datasets of multi-author writing style analysis in PAN 2025. The best results for each task— easy, medium, and hard—are highlighted in bold. As a first observation, we note that the best result for each task was achieved by a different team. This is consistent with last year's results. This year, the best result for the easy dataset was achieved by Boriceanu et al. [18] The best result for the medium dataset was achieved by Kaichuan Lin et al. [26] with an F_1 of 0.825, and the best result for the difficult dataset was achieved by Xiaocan Lin et al. [16] with an F_1 of 0.830.

Comparing these results with those of last year's edition of the task, it can be seen that a very similar level of accuracy was achieved at all three difficulty levels. Last year, the best results for the easy, medium, and hard datasets were 0.991, 0.887, and 0.863, respectively. This is despite the more difficult task this year, as style changes can now occur between individual sentences within a document, whereas last year this was only the case between paragraphs.

As in previous years, we again examined the accuracy of each team for documents with different numbers of authors. The results are shown in Figure 1 for the medium and the hard dataset. In the medium dataset, most of the submitted approaches show a peak in recognition accuracy for documents written by three authors, with accuracy decreasing for documents with four authors. The hard dataset shows a different picture: here, most approaches show a continuous increase in accuracy as the number of authors increases.

Finally, this year we also examined the computing costs of the individual approaches submitted for the first time. The results of this analysis for the hard dataset are shown in Table 2. The teams "hellojie" and "team-of-bf" are missing from this analysis because the required performance metrics could not be obtained. As can be seen from the data, the approaches with the highest runtime and computing costs are not necessarily those with the best accuracy. The runtime of the winning approach for the difficult dataset, "wqd," is average among all approaches.

5. Conclusion

In the 2025 edition of the multi-author writing style analysis task at PAN, we again asked participants to identify the positions in a text where the writing style changes. In this year's task, these style changes occurred at the sentence level, making the task more difficult than last year. A total of twelve software submissions and notebooks were submitted. Despite the higher level of difficulty of the task, the participants' results were roughly on par with last year's.

Table 2
Performance metrics for the approaches that were run on the hard data set. For CPU, RAM, GPU, VRAM, the values are the maximum utilization / usage during its execution.

Team	Team name	Hard F_1	Time (ms)	CPU	RAM	GPU	VRAM
Lin et al. [16]	team wqd	0.830	504,438	232	22020	55	1830
Lin et al. [26]	team xxsu-team	0.829	538,039	20	19660	49	5433
Boriceanu and Băltoiu [18]	team stylospies	0.791	6,190,189	166	22020	50	586
Hosseinbeigi and Mehrani [20]	team team-tmu	0.792	15,441,898	20	22020	8948	16395
Schmidt et al. [22]	team better-call-claude	0.731	50,341	232	22020	89	3300
Księżniak et al. [24]	team openfact	0.752	233,612	232	19660	96	3265
Bölöni-Turgut et al. [25]	team cornell-1	0.698	529,262	10	19660	98	4896
Rohra et al. [27]	team batatavada-pict	0.667	504,585	200	20709	73	808
Meier et al. [19]	team hhu	0.642	2,135,782	20	22020	94	2728
Alsheddi and Menai [17]	team ksu	0.467	656,520	144	22020	95	1691

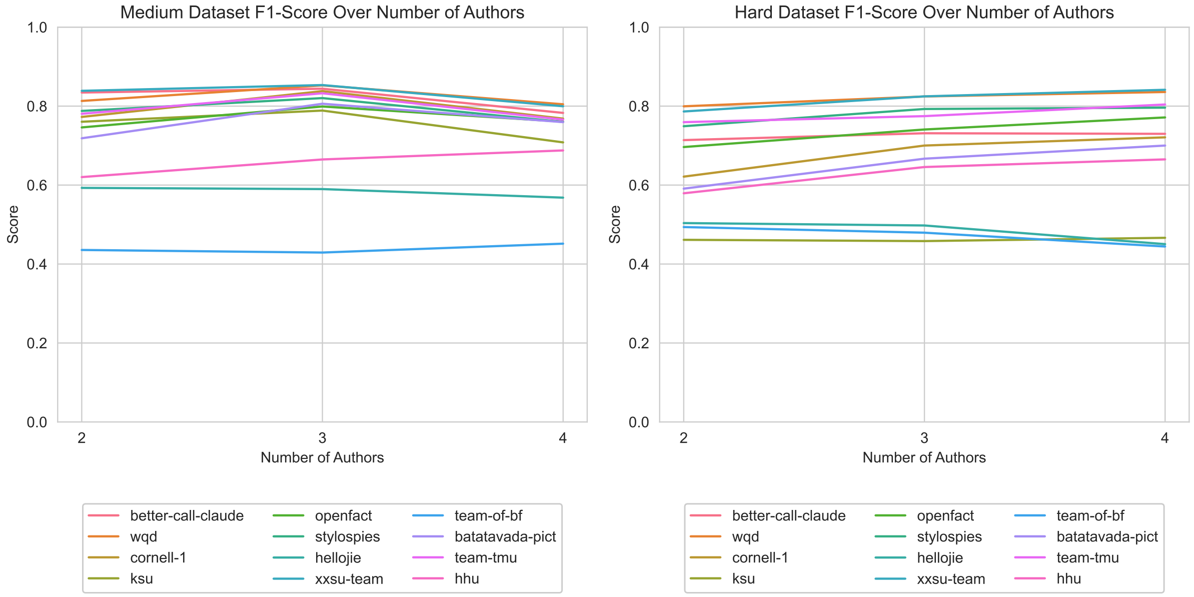

Figure 1: F_1 scores per team for the medium (left) and hard (right) datasets, depending on the number of authors per document.

Declaration on Generative AI

The authors used ChatGPT and DeepL for grammar and spelling checks as well as for paraphrasing. After using these services, the authors reviewed the content and revised it as necessary.

References

[1] E. Stamatatos, M. Tschuggnall, B. Verhoeven, W. Daelemans, G. Specht, B. Stein, M. Potthast, Clustering by Authorship Within and Across Documents, in: Working Notes Papers of the CLEF 2016 Evaluation Labs, CEUR Workshop Proceedings, CLEF and CEUR-WS.org, 2016. URL: http://ceur-ws.org/Vol-1609/.

[2] M. Kestemont, M. Tschuggnall, E. Stamatatos, W. Daelemans, G. Specht, B. Stein, M. Potthast, Overview of the Author Identification Task at PAN-2018: Cross-domain Authorship Attribution and Style Change Detection, in: L. Cappellato, N. Ferro, J.-Y. Nie, L. Soulier (Eds.), Working Notes Papers of the CLEF 2018 Evaluation Labs, volume 2125 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2018. URL: https://ceur-ws.org/Vol-2125/invited_paper_2.pdf.

[3] M. Tschuggnall, E. Stamatatos, B. Verhoeven, W. Daelemans, G. Specht, B. Stein, M. Potthast, Overview of the Author Identification Task at PAN 2017: Style Breach Detection and Author Clustering, in: L. Cappellato, N. Ferro, L. Goeuriot, T. Mandl (Eds.), Working Notes Papers of the CLEF 2017 Evaluation Labs, volume 1866 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2017. URL: http://ceur-ws.org/Vol-1866/.

[4] E. Zangerle, M. Tschuggnall, G. Specht, M. Potthast, B. Stein, Overview of the Style Change Detection Task at PAN 2019, in: L. Cappellato, N. Ferro, D. Losada, H. Müller (Eds.), CLEF 2019 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2019. URL: http://ceur-ws.org/Vol-2380/paper_243.pdf.

[5] E. Zangerle, M. Mayerl, G. Specht, M. Potthast, B. Stein, Overview of the Style Change Detection Task at PAN 2020, in: L. Cappellato, C. Eickhoff, N. Ferro, A. Névéol (Eds.), CLEF 2020 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2020. URL: https://ceur-ws.org/Vol-2696/paper_256.pdf.

[6] E. Zangerle, M. Mayerl, , M. Potthast, B. Stein, Overview of the Style Change Detection Task at PAN 2021, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021, pp. 1760–1771. URL: https://ceur-ws.org/Vol-2936/paper-148.pdf.

[7] E. Zangerle, M. Mayerl, M. Potthast, B. Stein, Overview of the Style Change Detection Task at PAN 2022, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), CLEF 2022 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2022. URL: http://ceur-ws.org/Vol-3180/paper-186.pdf.

[8] E. Zangerle, M. Mayerl, M. Potthast, B. Stein, Overview of the Multi-Author Writing Style Analysis Task at PAN 2023, in: M. Aliannejadi, G. Faggioli, N. Ferro, M. Vlachos (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), volume 3497 of *CEUR Workshop Proceedings*, 2023, pp. 2513–2522. URL: https://ceur-ws.org/Vol-3497/paper-201.pdf.

[9] E. Zangerle, M. Mayerl, M. Potthast, B. Stein, Overview of the Multi-Author Writing Style Analysis Task at PAN 2024, in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2024.

[10] E. Stamatatos, Intrinsic Plagiarism Detection Using Character n-gram Profiles, in: B. Stein, P. Rosso, E. Stamatatos, M. Koppel, E. Agirre (Eds.), SEPLN 2009 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 09), Universidad Politécnica de Valencia and CEUR-WS.org, 2009, pp. 38–46. URL: http://ceur-ws.org/Vol-502.

[11] A. Iyer, S. Vosoughi, Style Change Detection Using BERT—Notebook for PAN at CLEF 2020, in: L. Cappellato, C. Eickhoff, N. Ferro, A. Névéol (Eds.), CLEF 2020 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2020. URL: http://ceur-ws.org/Vol-2696/.

[12] Z. Zhang, Z. Han, L. Kong, X. Miao, Z. Peng, J. Zeng, H. Cao, J. Zhang, Z. Xiao, X. Peng, Style Change Detection Based On Writing Style Similarity—Notebook for PAN at CLEF 2021, in: G. Faggioli, N. Ferro, A. Joly, M. Maistro, F. Piroi (Eds.), CLEF 2021 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2021. URL: http://ceur-ws.org/Vol-2936/paper-198.pdf.

[13] T.-M. Lin, C.-Y. Chen, Y.-W. Tzeng, L.-H. Lee, Ensemble Pre-trained Transformer Models for Writing Style Change Detection, in: G. Faggioli, N. Ferro, A. Hanbury, M. Potthast (Eds.), CLEF 2022 Labs and Workshops, Notebook Papers, CEUR-WS.org, 2022. URL: http://ceur-ws.org/Vol-3180/paper-210.pdf.

[14] 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.

[15] J. Lv, Y. Yi, H. Qi, Team Fosu-stu at PAN: Supervised fine-tuning of large language models for Multi Author Writing Style Analysis, in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2024.

[16] X. Lin, Z. Han, C. Liu, X. Duan, 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, CEUR-WS.org, 2025.

[17] A. Alsheddi, M. El Bachir Menai, Style Change Detection in Multi-authored English Texts Based on Graph Convolutional Networks, in: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2025.

[18] I. Boriceanu, A. Băltoiu, Style Change Detection Using Graph and Structural-Linguistic Features for Multi-Author Writing Analysis, in: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2025.

[19] 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: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2025.

[20] S. B. Hosseinbeigi, A. Mehrani, Team TMU at PAN 2025: An Ensemble of Fine-Tuned LaBSE and Siamese Neural Network for Multi-Author Writing Style Analysis, in: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2025.

[21] D. Chen, J. Li, H. Qi, Llama-3 with 4-bit Quantization and IA^3 Tuning for Multi-Author Writing Style Analysis, in: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2025.

[22] G. Schmidt, J. Römisch, M. Halchynska, S. Gorovaia, I. Yamshchikov, better_call_claude: Sequential Style Shift Model for Fine-Grained Multi-Author Style Change Detection, in: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2025.

[23] B. Liu, L. Yang, H. Qi, Integrating Adversarial-Contrastive Learning and Large Language Model for Multi-Author Writing Style Analysis, in: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2025.

[24] E. Księżniak, K. Węcel, M. Sawiński, OpenFact at PAN 2025: Punctuation-Guided Pretraining for Sentence-Level Style Change Detection, in: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2025.

[25] D. Boloni-Turgut, D. Verma, C. Cardie, Team cornell-1 at PAN: Ensembling Fine-Tuned Transformer Models for Writing Style Analysis, in: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2025.

[26] 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, CEUR-WS.org, 2025.

[27] H. Rohra, N. Shah, S. Sonawane, Team BatataVada at PAN: Sentence-Level Style Change Detection with RoBERTa for Multi-Author Writing Style Analysis, in: Working Notes of CLEF 2025 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2025.