Team “better_call_claude”: Style Change Detection using a Sequential Sentence Pair Classifier*

Notebook for the PAN Lab at CLEF 2025

Gleb Schmidt[1,*,†], Johannes Römisch[2,†], Mariia Halchynska[2,†], Svetlana Gorovaia[3,*,†] and Ivan P. Yamshchikov[2]

[1]Humanities Lab, Faculaty of Arts, Radboud University, Houtlaan 4, 6525 XZ, Nijmegen, Netherlands
[2]Center for Artificial Intelligence, Technical University of Applied Sciences Würzburg-Schweinfurt, Münzstraße 12, 97070, Würzburg, Germany
[3]LEYA Lab, School of Computer Science, Physics and Technology, HSE University, 6, 25th Liniya, Vasilievsky Ostrov, 199004, St Petersburg, Russia

Abstract

Style change detection—identifying the points in a document where writing style shifts—remains one of the most important and challenging problems in computational authorship analysis. At PAN 2025, the shared task challenges participants to detect style switches at the most fine-grained level: individual sentences. The task spans three datasets, each designed with controlled and increasing thematic variety within documents. We propose to address this problem by modeling the content of each problem instance—that is, a series of sentences—as a whole, using a Sequential Sentence Pair Classifier (SSPC). The architecture leverages a pre-trained language model (PLM) to obtain representations of individual sentences, which are then fed into a bidirectional LSTM (BiLSTM) to contextualize them within the document. The BiLSTM-produced vectors of adjacent sentences are concatenated and passed to a multi-layer perceptron for prediction per adjacency. Building on the work of previous PAN participants classical text segmentation, the approach is relatively conservative and lightweight. Nevertheless, it proves effective in leveraging contextual information and addressing what is arguably the most challenging aspect of this year’s shared task: the notorious problem of “stylistically shallow”, short sentences that are prevalent in the proposed benchmark data. Evaluated on the official PAN 2025 test datasets, the model achieves strong macro-F1 scores of 0.923, 0.828, and 0.724 on the EASY, MEDIUM, and HARD data, respectively, outperforming not only the official random baselines but also a much more challenging one: `claude-3.7-sonnet`’s zero-shot performance.

Keywords

Style Change Detection, Text Segmentation, Sequence Labeling, BiLSTM, Large Language Models, Pre-Trained Language Models,

1. Introduction

Detecting changes in writing style—in other words, identifying places within a document where stylistic signal changes—is a form of authorship analysis that, perhaps alongside authorship verification, holds the greatest potential for applications beyond industrial contexts, particularly in humanities research. Given that our contemporary concept of individual authorship—let alone formal definitions of intellectual property and copyright—is a relatively recent development, a substantial part of human written culture goes back to periods when, broadly speaking, “collaborative writing” (actual co-authorship, extensive reuse, interpolation to mention but a few of its possible forms), was not only common—as it remains today—but was often regarded as a way of declaring ones belonging to a tradition, and therefore valued even more highly than original composition.

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

*You can use this document as the template for preparing your publication. We recommend using the latest version of the ceurart style.

*Corresponding author.

†These authors contributed equally.

gleb.schmidt@ru.nl (G. Schmidt); johannes.roemisch@study.thws.de (J. Römisch); mariia.halchynska@study.thws.de (M. Halchynska); sgorovaya@hse.ru (S. Gorovaia); ivan.yamshchikov@thws.de (I. P. Yamshchikov)

Nonetheless, the exploration of such "mixed-authorship texts" is typically hindered precisely by the uncertainties surrounding their authorial structure, which creates notorious contextualization problems and subsequently puts strict limitations on the interpretation of such texts.

Recent studies have shown that computational methods can offer valuable insights into mixed authorship at the level of entire corpora [1] or major subdivisions of individual works [2, 3, 4]. However, this level of granularity may often be insufficient for solving research questions that require a more fine-grained diarization—at the paragraph or even sentence level [5].

In this context, the decade-long effort of PAN organizers to stimulate research in this direction deserves special recognition. In various forms, the style change detection task has consistently been a part of the competition's program since 2016, making the participants' work notes and traditional overviews published in the aftermath of these events an invaluable source of methodological insight [6, 7]. Echoing the field's growing theoretical sophistication, for almost a decade the PAN workshops have been offering increasingly complex benchmark data and task formulations, encouraging participants to push the boundaries of achievable.

2. Related Work

2.1. Style Change Detection at PAN

Since its first inclusion in the PAN program in 2016, the style change detection task has appeared in various formulations. However, most of the factors contributing to its complexity were already present in the first two editions—namely, the required level of granularity for document analysis, the uncertainty regarding the number of style changes or contributing authors, and the need to segment the document. The most recent development of the task introduced only one additional—though important—dimension: controlled topic diversity in the data.

2.1.1. 2016—2022: In Search of Task's Score

At PAN 2016, the task was framed as an *authorship diarization* problem, closely related to the traditional intrinsic plagiarism detection explored during the early years of the competition's history [8, 9]. Participants faced three subtasks, each highlighting different challenges that were to become recurring focal points of the task in the following years. The first subtask assumed a major author and required identifying segments written by secondary authors. In the second subtask, the number of authors was provided, and participants were required to cluster document segments by authorship. The third and most challenging subtask involved building authorial clusters without any prior knowledge of the number of authors.

Operating at the sentence level, both proposed methods relied on traditional stylometric features, which were then processed using techniques typical of intrinsic plagiarism detection—namely, threshold- or Gaussian HMM-based outlier detection [10], and clustering [11]. These approaches, however, failed to achieve sufficient performance at such a fine level of granularity.

The poor performance led to a redefinition of the task as a style breach detection problem at PAN 2017. Instead of complete clustering a document's segments by authorship, participants were asked to predict whether a document was written by multiple authors and, if so, to identify the points at which the writing style changes [12].

The submitted approaches again centered on distance measures and outlier detection applied to sentences as well as actual or artificially constructed paragraphs, represented using either conventional stylometric features [13, 14] or neural sentence embeddings [15]. Despite the relative improvement over PAN 2016, the results of PAN 2017 confirmed that style-based document decomposition remained marginally beyond the state of the art at that time.

Therefore, for PAN 2018, the task was redefined once again, framing the problem as a document-level binary classification task. This invited participants to explore how stylistic inconsistency could be detected across an entire text [16]. The submissions reflected both conventional feature engineering

combined with rule-based or classical machine learning approaches [17, 18, 19] and early applications of deep learning [20, 21]. Ensembling multiple classifiers operating on diverse feature spaces—each capturing different aspects of language—not only proved reliable [19, 18], but also yielded the winning result [18]. The core of the classifier proposed by [21] was a CNN designed to capture patterns of character bigrams in groups of varying length.

The submission by [20] deserves special attention not only because it scored second, but also because it anticipated developments in recent authorship analysis research and served as a distant source of inspiration for the approach proposed below. Instead of relying on traditional feature spaces—where lexical components took center stage at the time—[20] focused exclusively on the dependency trees of sentences. The expressiveness of this feature space has recently been confirmed in a series of authorship attribution studies [22, 23, 24, 25]. [20] define and extract what they call a Parse Tree Feature (PTF)—a path from the root to any given word—and use it to represent each sentence as a sequence of its words' PTFs, and each document as a sequence of sentence representations. Subsequently, both the original and reversed versions of the document are fed into an LSTM with an additional sentence-level attention mechanism, which contextualizes each sentence based on rich syntactic information across the entire problem. Finally, the similarity between the original and reversed representations is computed and used as the basis for prediction.

Responding to performance boost observed at PAN 2018 on a simplified version of the problem, the organizers of PAN 2019 increased the task's complexity once again by adding the objective of predicting the number of authors per document [26]. To address this task, [27] employed a combination of clustering techniques based on representations of balanced-size text chunks obtained using the 50 most frequent words (MFW). Relying on a multi-layer perceptron operating on tf-idf-weighted word unigrams to detect style changes within a document, Zuo et al. subsequently used an extension of [18]'s feature extraction procedure to represent document paragraphs. They then applied an ensemble model comprising two clustering methods and a multi-layer perceptron to predict the number of style changes.

The Style Change Detection task at PAN 2020 was marked by two important shifts. First, after a significant departure from its originally intended scope during PAN 2018—2019, "the task was steered back into its original direction" [29, 1]. The segmentation component was reintroduced, and in addition to the document-level prediction of multi-authorship, participants were required to identify the positions where paragraph-level style changes occur. Second, for the first time, a solution based on pre-trained transformers was employed to address the task [30], significantly outperforming clustering-based approaches—the B_0-maximal used by [31] and a modified version of [27], which, however, remained undocumented.

PAN 2021 reintroduced yet another element of the task's original scope that had previously been set aside as too complex: grouping of text segments by authorship within documents. The shared task presented arguably the most complete formulation of the problem, comprising three separate questions: (1) whether the text was written by multiple authors; (2) where between paragraphs the writing style changes; and (3) which author each paragraph belongs to [32]. Although the competition saw an increasing reliance on pre-trained transformers, it was marked by a wide diversity of methods. [33] proposed the highest-scoring approach for Tasks 2 and 3, using a similarity measure extracted from paragraphs with a pre-trained BERT model. They approached all tasks simultaneously, first solving Tasks 2 and 3 in an authorship verification fashion. Each paragraph was compared with all preceding ones, and a new authorial class was assigned whenever a paragraph could not be classified as written by the same author as any of the previous ones. This information was subsequently used to solve the remaining tasks. The approach by [34] excelled at Task 1. It combined sentence features extracted using BERT and aggregated per paragraphs with the set of stylometric features proposed by [18]. The tasks were solved by stacking two feature spaces and feeding them to an ensemble of four classifiers. [35] decomposed the tasks into a series of authorship verification problems and solved them adapting the method proposed by [36]. Other works operated over various selections of stylometric features and used LSTM-powered model [37] and Siamese architecture [38] respectively.

Three subtasks of PAN 2022 challenged participants with both segmentation task and granularity level. Task 1 required finding the only style shift in a document co-authored by two persons. In Tasks 2

and 3, it was necessary to find style changes in a text written by two or more authors with switches occurring at only paragraph or paragraph and sentence levels. Despite clear prevalence of pre-trained transformer-based approaches, submissions exclusively working with manually-engineered feature spaces and traditional classification or clustering approaches [39, 40] or combining hand-picked features with those extracted using pre-trained models were submitted [41]. One of the submissions downright "hacked" the task by accessing extrinsic information online and yielding a nearly perfect result [42]. Most submissions, however, explored different PLM-based architectures. The overall best score was achieved by [43] who obtained predictions for pairs of sentences or paragraphs by ensembling classifiers based on BERT, RoBERTa, and ALBERT. [44] classified pairs of sentence or paragraph representations obtained applying one-dimensional convolution and max pooling to BERT output. [45] used a prompt-based approach fine-tuning a BERT model with masked language modeling objective to predict special tokens such <DIFFERENT> or <SAME> in a dynamically-constructed sequence: "They are the <MASK> writing style: Para1 and Para2". [46] trained three different transformer models to address each subtask. [47] used LSTM, convolution, and max pooling over BERT-based word representations.

2.1.2. 2023—2024: Strengthening Connections to Real-World Scenarios

Two past shared tasks are both characterized by explicit intention to put the theoretical problem closer to real-world scenarios and addressed the problem of possible topic consistency within the documents introducing controlled levels of thematic homogeneity in benchmark data challenging participants with development of methods less dependent on thematic signal.

The solutions submitted to PAN 2023 demonstrated relative difficulty of this setting. Whereas most submission achieved F1 score of more than 90% and 80% on EASY and MEDIUM tasks where writing style change could coincide with thematic shift, the performance on single-topic HARD dataset was significantly lower.

The year was marked by further expansion of the PLMs' use, although one solution focusing on traditional stylometry was also submitted [48]. One of the important tendencies that year was a broad diversity of ways in which PLMs's linguistic knowledge was integrated into the solutions. [49, 50] made recourse to contrastive learning in former case combining it with a prompt-based approach that excelled on the HARD dataset. [51] solved the task as inference problem concatenating paragraphs and predicting special tokens <ENTAILMENT> or <CONTRADICTION>. [52] pre-trained a custom model serving as a basis for classifier, while [53] achieved highest scores on EASY and MEDIUM data using an ensemble of several PLMs to predict binary labels for concatenated pairs of adjacent paragraphs.

The following year's shared task on multi-author style analysis retained the same definition and structure of benchmark data as 2023 [54], continuing the focus on paragraph-level style change detection with varying levels of topical homogeneity. Contrastive learning techniques and ensemble architectures based on large PLMs took an even more prominent role. As a result, overall performance improved and the gap between multi-topic and single-topic document scenarios narrowed. The best submission achieved an impressive F1 around 86% on HARD dataset. Notably, purely traditional stylometry approaches virtually disappeared in 2024, as nearly all top methods relied on fine-tuned transformer models (often in combination or with specialized training objectives) to detect writing style changes.

2.1.3. Generated Text Detection in Human-AI Collaborative Hybrid Texts

The surge of Generative AI has given rise to a new field of application for methods conceptually related to style change detection—authorship analysis in hybrid documents, i.e., texts co-written by humans and AI. Zeng et al. investigated the detection of AI-generated passages within human–AI collaborative texts, highlighting unique challenges of this setting: frequent author switches, obfuscation by post-editing, and the limited availability of stylistic cues in short segments. The suggested approach—a two-step segmentation and classification method augmented with modern transformers and contrastive techniques—builds directly on the foundations laid by style change detection research.

This hybrid document segmentation problem was also the focus of the ALTA 2024 shared task

[56], which required participants to identify AI-generated sentences in news articles. The competition demonstrated a clear methodological convergence with style change detection research at PAN.

3. Methodology

3.1. Challenges of Style Change Detection task at PAN 2025

At first glance, this year's task may appear less challenging. On the one hand, it does not require the explicit grouping of text segments by authorship. On the other, sentence-level granularity is by no means new and has been successfully addressed in previous editions. Yet, as the organizers note, the benchmark data was designed to more accurately replicate real-world scenarios, which is why the level of thematic coherence within each document was meticulously controlled [6].

Whereas the documents in the EASY dataset always cover multiple topics, the MEDIUM and HARD datasets exhibit limited or no thematic diversity, respectively. Therefore, while in the first case solutions may rely on thematic clues as potential indicators of style change, the latter two—each to a different extent—force participants to rely more heavily on detecting subtle style changes rather than topic variations [6, 438].

Further challenges of this year's shared task become evident in preliminary exploratory analysis of the benchmark data, which reveals several peculiarities of the data that significantly amplify the difficulty of the task (see Appendix A):

- Relatively short sentence length.
- A substantial portion—over 10%—of all sentences are *exact duplicates*, with some occurring more than 3,000 times[1].
- Hundreds of sequences only contain punctuation marks, but are placed on separate lines and thus formally treated as separate sentences by the compilers of the data.

Consequently, each individual sentence may not provide sufficient information for identifying author's fingerprint and making a reliable decision. A pair of identical or nearly identical one- or two-word sentences—not only fairly common in Internet communication in general and also abundantly present in the data—may be entirely style- or even content-neutral, i.e., provide no reliable clues whatsoever.

3.2. Core Intuition

To address the notorious problem of such "shallow sentences", we pivoted our approach around the idea of incorporating into the decision-making process the one thing that even the most minimalist one-word sentence always has—its context, or, in simpler terms, its position *within the problem*. Therefore, we designed a BiLSTM-powered solution intended to model a problem *as a whole* and capture positional dependencies between the document's sentences treated as atomic units that are organized into stylistically cohesive segments.

Our inspiration comes from late 2010s work on text segmentation and learning cohesion breaks. [57] demonstrated efficiency of bidirectional RNN trained on positive and negative examples of cohesive text in learning breaks in speech transcriptions. Further theoretical step was made by [58] who presented text topic segmentation task as a supervised, specifically—sequence labeling, problem and employed to a BiLSTM powered architecture operating over sentence embeddings to implement this approach. [59]'s system, SegBot, achieved reliable performance on segmentation task at sentence and Elementary Discourse Unit (EDU) level. Improving and expanding the method, [60] implemented a system segmenting texts into thematically coherent sections and assigning topic labels. Glavaš and

[1]The most frequently repeated sentences include moderation messages (e.g., "Debate/discuss/argue the merits of ideas, don't attack people.", "r/politics is currently accepting new moderator applications.") and automatic notifications (e.g., "Personal insults, shill or troll accusations, hate speech, any suggestion or support of harm, violence, or death, and other rule violations can result in a permanent ban.", "I am a bot, and this action was performed automatically.")

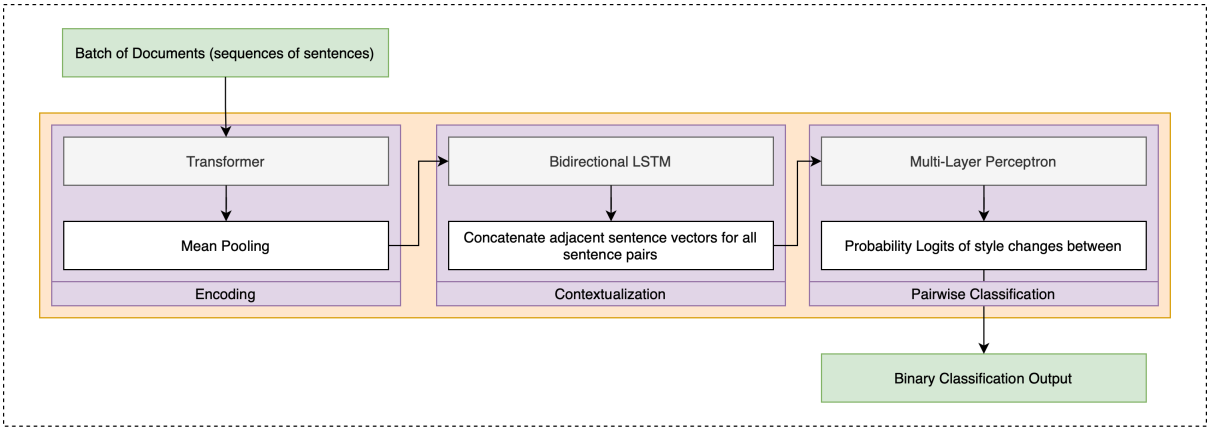

Figure 1: Sequential Sentence Pair Classifier

Somasundaran and Lo et al. implemented similar contextualization approaches employing two-level transformers.

The approach also has predecessors among PAN participants. Hosseinia and Mukherjee treated representations of problem's sentences with syntactic features as atomic units and explored their sequential dependencies using an LSTM. More recently, BiLSTM appeared several times as a steps in extraction of sentence or paragraph representation from PLMs [37, 47].

3.3. Architecture

To implement the idea, we opted for a lightweight solution. A problem—a sequence of sentences—is considered a single sample, and its sentences are encoded using a PLM. Applying straightforward mean pooling to token embeddings, fixed-length representations of the problem's sentences are obtained. Then, to capture the inter-sentence contextual clues, the sequence of problem's sentence vectors is feed into a BiLSTM. This layer outputs context-aware sentence representations enriching raw mean-pooled vectors with information from the sentences across the entire problem. Subsequently, by concatenating each sentence vector with adjacent one across feature dimensions, representations of pairs of adjacent sentences are constructed. These are then passed through a multi-layer perceptron (MLP) classifier that outputs logits corresponding to the probability of a style change between each sentence pair (see Figure 1).

This design enables the model to leverage both the semantic richness of the fine-tuned backbone PLM and the sequential structure of the problem, resulting in robust style change detection performance across documents of varying lengths and complexities. The submitted implementation code is available on GitHub[2].

3.3.1. Base Transformer

Different pre-trained models were tested (see Table 1), but `StyleDistance/styledistance`, presented by [63], was retained and submitted for evaluation on the test data.

4. Training

4.1. Data Augmentation

To obtain more training data, we used three datasets from PAN 2024. All sentence transitions within a single paragraph were labeled as not representing a style change. The first sentence of each new paragraph was labeled as 1, indicating a style change.

[2]https://github.com/glsch/better-call-claude_pan25-multi-author-style-analysis

Table 1
F1 scores for each model, task, and dataset configuration.

Model	2025 data			2025 augmented with 2024 data		
	Easy	Medium	Hard	Easy	Medium	Hard
RoBERTa-base [64]	**0.929**	0.764	0.613	**0.930**	0.787	0.656
XLM-RoBERTa-base [65]	0.927	0.779	0.540	0.929	0.779	0.658
all-MiniLM-L6-v2 [66]	0.878	0.763	0.607	0.906	0.777	0.654
StyleDistance [63]	0.891	**0.780**	**0.629**	0.922	**0.828**	**0.723**

A single model for all subtasks was trained on the most complete training data: the three sentence-level datasets from 2025 and all three paragraph-turned-sentence-level datasets from 2024.

4.2. Training

The Sequential Sentence Pair Classifier was implemented in PyTorch Lightning, allowing for seamless experimentation with both the architecture and hyperparameters.

Table 2
Model hyperparameters.

Hyperparameter	
Base transformer	`StyleDistance/styledistance`
Base transformer frozen	True
Pooling	Mean
BiLSTM layers	5
BiLSTM dropout	0.2
MLP	A three-layer feedforward network with linear layers, GELU activations, and dropout
MLP dropout	0.2
Batch	4
Learning rate	0.0005
Minimal learning rate	0.00005
Scheduler	cosine
Training steps	30000
Warmup steps	2600
Loss	Binary Cross-Entropy

Table 3
Performance of Claude and Sequential Sentence Pair Classifier on the official PAN 2025 test data.

Model	Dataset	F1 (macro)
Claude	hard	0.661
	medium	**0.818**
	easy	0.856
Sequential Sentence Pair Classifier	hard	**0.731**
	medium	0.815
	easy	**0.929**

5. Baselines

At PAN 2024, three baselines were used: RANDOM, PREDICT 1, and PREDICT 0. While using all these baselines, we decided on another—and more challenging—one, LLM, zero-shot predictions by

a best-performing LLM `claude-3.7-sonnet` prompted with the so-called linguistically informed prompts (LIP) [67]. For the detailed description of the baseline setup we address the reader to [68].

6. Results

The following results were obtained on the official PAN 2025 test data (Table 3).

The gradual decline in performance from the easy to the hard tasks reflects the increasing difficulty of identifying style changes. Overall, our model outperforms zero-shot large language model predictions on both the easy and hard tasks and falls short by only a fraction of a percentage point on the medium task. At the same time, the proposed solution is lightweight and does not require much computational resources.

7. Discussion

Several observations stem from our approach to the PAN 2025 style change detection task.

First, while our model is effective, its strength lies in exploiting the macrostructure of the problem and inter-sentence contextual patterns—particularly the sequential order and distribution of sentences—rather than in isolating purely stylistic signals. This reflects a broader shift from traditional stylometric analysis, which typically assumes topical uniformity and relies on intrinsic features like syntax and lexical choice. This raises concerns about generalizability.

Future iterations of the task could focus more on isolation of stylistic signal by reducing contextual cues—e.g., further controlling topic coherence or randomizing sentence order—to more rigorously test a model's ability to capture intrinsic authorial style.

Second, the strong zero-shot performance of Claude draws attention to the growing impact of LLMs in authorship analysis. LLMs, with their vast pretraining and generalization capabilities, can recognize both contextual and stylistic patterns with little to no task-specific adjustment. Future PAN tasks might clearly separate evaluation tracks that allow external LLM calls from those that do not even for otherwise AI-unrelated tasks.

Conclusion

Despite the promising results of the proposed model, it has several limitations. First, the padding strategy required for batch processing may hinder scalability and efficiency when applied to much longer texts or inputs with highly variable lengths. Second, while the BiLSTM used for contextualization has proven effective, it may not be the optimal architecture for capturing complex dependencies, particularly in longer input sequences. More sophisticated architectures—such as those proposed by [61, 62]—could potentially yield better results.

Finally, our reliance on a frozen pre-trained encoder may limit the model's adaptability to domain-specific nuances. Fine-tuning the encoder or incorporating domain-specific data and training strategies could further improve performance.

Declaration on Generative AI

During the preparation of this work, the author(s) used Claude 4 and GPT-4 (gpt-3.5/gpt-4) to perform grammar/ spelling checks and edit text for clarity. Additionally, the author(s) used Perplexity's research tools to double-check relevant literature and contributions related to the topic. All content was reviewed and edited by the author(s), who take full responsibility for the final publication.

References

[1] T. Clérice, A. Glaise, Twenty-One* Pseudo-Chrysostoms and more: authorship verification in the patristic world, in: Computational Humanities Research Conference 2023, Proceedings of the Computational Humanities Research Conference 2022, Paris, France, 2023. URL: https://inria.hal.science/hal-04211176.

[2] F. Cafiero, J.-B. Camps, 'Psyché'as a Rosetta Stone? Assessing Collaborative Authorship in the French 17th Century Theatre, Proceedings http://ceur-ws. org ISSN 1613 (2021) 0073.

[3] P. Plecháč, Relative contributions of Shakespeare and Fletcher in Henry VIII: An analysis based on most frequent words and most frequent rhythmic patterns, Digital Scholarship in the Humanities 36 (2021) 430–438. Publisher: Oxford University Press.

[4] G. Schmidt, V. Vybornaya, I. P. Yamshchikov, Fine-Tuning Pre-Trained Language Models for Authorship Attribution of the Pseudo-Dionysian Ars Rhetorica, Aarhus, 2024.

[5] M. Eder, Rolling stylometry, Digital Scholarship in the Humanities 31 (2016) 457–469. Publisher: Oxford University Press.

[6] 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: Generative AI Detection, Multilingual Text Detoxification, Multi-author Writing Style Analysis, and Generative Plagiarism Detection, in: C. Hauff, C. Macdonald, D. Jannach, G. Kazai, F. M. Nardini, F. Pinelli, F. Silvestri, N. Tonellotto (Eds.), Advances in Information Retrieval, Springer Nature Switzerland, Cham, 2025, pp. 434–441. URL: https://doi.org/10.1007/978-3-031-88720-8_64. doi:10.1007/978-3-031-88720-8_64.

[7] E. Zangerle, M. Mayerl, M. Potthast, B. Stein, Multi-Author Writing Style Analysis 2025, 2025. URL: https://pan.webis.de/clef25/pan25-web/style-change-detection.html.

[8] 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/Balog, Krisztian [edit.]; et al., 2016, pp. 691–715.

[9] P. Rosso, F. Rangel, M. Potthast, E. Stamatatos, M. Tschuggnall, B. Stein, Overview of PAN'16: new challenges for authorship analysis: cross-genre profiling, clustering, diarization, and obfuscation, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction: 7th International Conference of the CLEF Association, CLEF 2016, Évora, Portugal, September 5-8, 2016, Proceedings 7, Springer, 2016, pp. 332–350.

[10] M. P. Kuznetsov, A. Motrenko, R. Kuznetsova, V. V. Strijov, Methods for Intrinsic Plagiarism Detection and Author Diarization., in: CLEF (Working notes), 2016, pp. 912–919.

[11] A. Sittar, H. R. Iqbal, R. M. A. Nawab, Author Diarization Using Cluster-Distance Approach., in: CLEF (Working Notes), 2016, pp. 1000–1007.

[12] 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: Working Notes Papers of the CLEF 2017 Evaluation Labs/Cappellato, Linda [edit.]; et al., 2017, pp. 1–22.

[13] D. Karas, M. Spiewak, P. Sobecki, OPI-JSA at CLEF 2017: Author Clustering and Style Breach Detection., in: CLEF (Working Notes), 2017.

[14] J. A. Khan, Style Breach Detection: An Unsupervised Detection Model., in: CLEF (Working Notes), 2017.

[15] K. Safin, R. Kuznetsova, Style Breach Detection with Neural Sentence Embeddings., in: CLEF (Working Notes), 2017.

[16] 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: Working Notes Papers of the CLEF 2018 Evaluation Labs. Avignon, France, September 10-14, 2018/Cappellato, Linda [edit.]; et al., 2018, pp. 1–25.

[17] J. A. Khan, A model for style change detection at a glance, volume 593, 2018, p. 113.

[18] D. Zlatkova, D. Kopev, K. Mitov, A. Atanasov, M. Hardalov, I. Koychev, P. Nakov, An ensemble-rich

multi-aspect approach for robust style change detection, in: CLEF 2018 Evaluation Labs and Workshop–Working Notes Papers, CEUR-WS. org, 2018, p. 3.
[19] K. Safin, A. Ogaltsov, Detecting a change of style using text statistics, Working Notes of CLEF (2018).
[20] M. Hosseinia, A. Mukherjee, A Parallel Hierarchical Attention Network for Style Change Detection, CLEF, 2018.
[21] N. Schaetti, Character-based Convolutional Neural Network for Style Change Detection: Notebook for PAN at CLEF 2018., in: CLEF (Working Notes), 2018.
[22] R. Gorman, Author identification of short texts using dependency treebanks without vocabulary, Digital Scholarship in the Humanities 35 (2020) 812–825. URL: https://academic.oup.com/dsh/article/35/4/812/5606771. doi:10.1093/llc/fqz070.
[23] R. Gorman, Universal Dependencies and Author Attribution of Short Texts with Syntax Alone., DHQ: Digital Humanities Quarterly 16 (2022).
[24] R. Gorman, Morphosyntactic Annotation in Literary Stylometry, Information 15 (2024) 211. URL: https://www.mdpi.com/2078-2489/15/4/211. doi:10.3390/info15040211.
[25] V. B. Gorman, R. J. Gorman, A morphosyntactic authorship attribution study of the speeches of Demosthenes and Apollodorus, The Journal of Hellenic Studies 144 (2024) 65–92. URL: https://www.cambridge.org/core/product/identifier/S0075426924000302/type/journal_article. doi:10.1017/S0075426924000302.
[26] E. Zangerle, M. Tschuggnall, G. Specht, B. Stein, M. Potthast, Overview of the Style Change Detection Task at PAN 2019, in: L. Cappellato, N. Ferro, D. E. Losada, H. Müller (Eds.), Working Notes Papers of the CLEF 2019 Evaluation Labs, volume 2380 of *CEUR Workshop Proceedings*, 2019. URL: https://ceur-ws.org/Vol-2380/paper_243.pdf, iSSN: 1613-0073.
[27] S. Nath, Style change detection by threshold based and window merge clustering methods., in: CLEF (Working Notes), 2019.
[28] C. Zuo, Y. Zhao, R. Banerjee, Style Change Detection with Feed-forward Neural Networks., CLEF (Working Notes) 93 (2019).
[29] E. Zangerle, M. Mayerl, G. Specht, B. Stein, M. Potthast, Overview of the Style Change Detection Task at PAN 2020, in: L. Cappellato, C. Eickhoff, N. Ferro, A. Névéol (Eds.), Working Notes Papers of the CLEF 2020 Evaluation Labs, volume 2696 of *CEUR Workshop Proceedings*, 2020. URL: https://ceur-ws.org/Vol-2696/paper_256.pdf, iSSN: 1613-0073.
[30] A. Iyer, S. Vosoughi, Style Change Detection Using BERT., CLEF (Working Notes) 93 (2020) 106.
[31] D. Castro-Castro, C. A. Rodríguez-Lozada, R. Muñoz, Mixed Style Feature Representation and B-maximal Clustering for Style Change Detection., in: CLEF (Working Notes), 2020.
[32] 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.), Working Notes Papers of the CLEF 2021 Evaluation Labs, volume 2936 of *CEUR Workshop Proceedings*, 2021. URL: https://ceur-ws.org/Vol-2936/paper-148.pdf, iSSN: 1613-0073.
[33] 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.
[34] E. Strøm, Multi-label Style Change Detection by Solving a Binary Classification Problem., in: CLEF (working notes), 2021, pp. 2146–2157.
[35] R. Singh, J. Weerasinghe, R. Greenstadt, Writing Style Change Detection on Multi-Author Documents., in: CLEF (Working Notes), 2021, pp. 2137–2145.
[36] J. Weerasinghe, R. Greenstadt, Feature vector difference based neural network and logistic regression models for authorship verification, in: CEUR workshop proceedings, volume 2695, 2020.
[37] R. Deibel, D. Löfflad, Style Change Detection on Real-World Data using an LSTM-powered Attribution Algorithm., in: CLEF (Working Notes), 2021, pp. 1899–1909.
[38] S. Nath, Style change detection using Siamese neural networks., in: CLEF (Working Notes), 2021,

pp. 2073–2082.
[39] S. Alshamasi, M. Menai, Ensemble-Based Clustering for Writing Style Change Detection in Multi-Authored Textual Documents, 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-187.pdf.
[40] H. A. F. Alvi, N. Alqahtani, Style Change Detection using Discourse Markers, 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-188.pdf.
[41] C. A. Rodríguez-Losada, D. Castro-Castro, Three Style Similarity: sentence-embedding, auxiliary words, punctuation, 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-218.pdf.
[42] L. Graner, P. Ranly, An Unorthodox Approach for Style Change Detection., in: CLEF (Working Notes), 2022, pp. 2455–2466.
[43] T.-M. Lin, C.-Y. Chen, Y.-W. Tzeng, L.-H. Lee, Ensemble Pre-trained Transformer Models for Writing Style Change Detection., in: CLEF (Working Notes), 2022, pp. 2565–2573.
[44] Q. Lao, L. Ma, W. Yang, Z. Yang, D. Yuan, Z. Tan, L. Liang, Style Change Detection Based On Bert And Conv1d, 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-208.pdf.
[45] Z. Zhang, Z. Han, L. Kong, Style Change Detection based on Prompt., in: CLEF (Working Notes), 2022, pp. 2753–2756.
[46] Z. Z. X. Jiang, M. Huang, Style Change Detection: Method Based On Pre-trained Model And Similarity Recognition, 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-205.pdf.
[47] L. Z. J. Zia, Z. Liua, Style Change Detection Based On Bi-LSTM And Bert, 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-234.pdf.
[48] G. Jacobo, V. Dehesa, D. Rojas, H. Gómez-Adorno, Authorship verification machine learning methods for Style Change Detection in texts, 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. 2652–2658. URL: https://ceur-ws.org/Vol-3497/paper-217.pdf.
[49] Z. Ye, C. Zhong, H. Qi, Y. Han, Supervised Contrastive Learning for Multi-Author Writing Style Analysis., in: CLEF (Working Notes), 2023, pp. 2817–2822.
[50] W. Chen, X. Liu, Contrastive Learning Approaches for Multi-Author Style Analysis, in: CLEF 2023 Working Notes, volume 3497, CEUR-WS, 2023.
[51] I. E. Kucukkaya, U. Sahin, C. Toraman, ARC-NLP at PAN 23: Transition-Focused Natural Language Inference for Writing Style Detection, 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. 2659–2668. URL: https://ceur-ws.org/Vol-3497/paper-218.pdf.
[52] M. Huang, Z. Huang, L. Kong, Encoded Classifier Using Knowledge Distillation 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. 2629–2634. URL: https://ceur-ws.org/Vol-3497/paper-214.pdf.
[53] A. Hashemi, W. Shi, Enhancing Writing Style Change Detection using Transformer-based Models and Data Augmentation., in: CLEF (Working Notes), 2023, pp. 2613–2621.
[54] 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. Herrera (Eds.), Working Notes Papers of the CLEF 2024 Evaluation Labs, CEUR-WS.org, 2024, pp. 2513–2522. URL: http://ceur-ws.org/Vol-3740/paper-222.pdf.
[55] Z. Zeng, S. Liu, L. Sha, Z. Li, K. Yang, S. Liu, D. Gašević, G. Chen, Detecting AI-Generated

Sentences in Human-AI Collaborative Hybrid Texts: Challenges, Strategies, and Insights, 2024. URL: http://arxiv.org/abs/2403.03506. doi:10.48550/arXiv.2403.03506, arXiv:2403.03506 [cs].
[56] D. Mollá, Q. Xu, Z. Zeng, Z. Li, Overview of the 2024 alta shared task: Detect automatic ai-generated sentences for human-ai hybrid articles, arXiv preprint arXiv:2412.17848 (2024).
[57] I. Sehikh, D. Fohr, I. Illina, Topic segmentation in ASR transcripts using bidirectional RNNs for change detection, in: 2017 IEEE automatic speech recognition and understanding workshop (ASRU), IEEE, 2017, pp. 512–518.
[58] O. Koshorek, A. Cohen, N. Mor, M. Rotman, J. Berant, Text Segmentation as a Supervised Learning Task, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 469–473. URL: http://aclweb.org/anthology/N18-2075. doi:10.18653/v1/N18-2075.
[59] J. Li, A. Sun, S. R. Joty, SegBot: A Generic Neural Text Segmentation Model with Pointer Network., in: IJCAI, 2018, pp. 4166–4172.
[60] S. Arnold, R. Schneider, P. Cudré-Mauroux, F. A. Gers, A. Löser, SECTOR: A Neural Model for Coherent Topic Segmentation and Classification, Transactions of the Association for Computational Linguistics 7 (2019) 169–184. URL: https://direct.mit.edu/tacl/article/43514. doi:10.1162/tacl_a_00261.
[61] G. Glavaš, S. Somasundaran, Two-Level Transformer and Auxiliary Coherence Modeling for Improved Text Segmentation, Proceedings of the AAAI Conference on Artificial Intelligence 34 (2020) 7797–7804. URL: https://ojs.aaai.org/index.php/AAAI/article/view/6284. doi:10.1609/aaai.v34i05.6284.
[62] 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, 2021. URL: http://arxiv.org/abs/2110.07160. doi:10.48550/arXiv.2110.07160, arXiv:2110.07160 [cs].
[63] 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, 2024. URL: https://arxiv.org/abs/2410.12757. doi:10.48550/ARXIV.2410.12757, version Number: 2.
[64] L. Zhuang, L. Wayne, S. Ya, Z. Jun, A Robustly Optimized BERT Pre-training Approach with Post-training, in: S. Li, M. Sun, Y. Liu, H. Wu, K. Liu, W. Che, S. He, G. Rao (Eds.), Proceedings of the 20th Chinese National Conference on Computational Linguistics, 2021, pp. 1218–1227. URL: https://aclanthology.org/2021.ccl-1.108/.
[65] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, V. Stoyanov, Unsupervised Cross-lingual Representation Learning at Scale, in: D. Jurafsky, J. Chai, N. Schluter, J. Tetreault (Eds.), Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 8440–8451. URL: https://aclanthology.org/2020.acl-main.747/. doi:10.18653/v1/2020.acl-main.747.
[66] N. Reimers, I. Gurevych, Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2019. URL: https://arxiv.org/abs/1908.10084.
[67] B. Huang, C. Chen, K. Shu, Can Large Language Models Identify Authorship?, 2024. URL: http://arxiv.org/abs/2403.08213. doi:10.48550/arXiv.2403.08213, arXiv:2403.08213 [cs].
[68] G. Schmidt, J. Römisch, I. Yamshchikov, S. Gorovaia, M. Halchynska, Better Call Claude: Can LLMs Detect Changes of Writing Style?, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), 2025.

A. Dataset Statistics

Tables 4 and 5 represent general data statistics and top-5 duplicated sentences in the data.

Table 4
Document statistics for different subsets of the dataset

Subset	Problems	Sentences	Avg words/sent.	Median words/sent.	Avg sent./doc	Median sent./doc
Easy	5100	63584	16.9 ± 16.4	13.0	12.5 ± 4.1	12.0
Medium	5100	76379	17.7 ± 11.9	15.0	15.0 ± 9.4	12.0
Hard	5100	66555	18.7 ± 12.0	17.0	13.0 ± 4.7	12.0
All Data	15300	206518	17.8 ± 13.5	15.0	13.5 ± 6.6	12.0

Table 5
Top-5 most frequent sentences.

Sentence	Count
In general, be courteous to others.	3296
Debate/discuss/argue the merits of ideas, don't attack people.	3296
Personal insults, shill or troll accusations, hate speech, any suggestion or support of harm, violence, or death, and other rule violations can result in a permanent ban.	3296
For those who have questions regarding any media outlets being posted on this subreddit, please click to review our details as to our approved domains list and outlet criteria.	2953
r/politics is currently accepting new moderator applications.	2172

B. Online Resources

- GitHub