Sentence-Level Style Change Detection with RoBERTa for Multi-Author Writing Style Analysis

Notebook for the PAN Lab at CLEF 2025

Harjas Rohra[1,*,†], Nirbhay Shah[1,†] and Sheetal Sonawane[1,†]

[1]*Pune Institute of Computer Technology, Pune*

Abstract

Style change detection aims to identify points within a document where authorship shifts occur, which is crucial for tasks like plagiarism detection, authorship verification, and writing support. This submission addresses the intrinsic style change detection task from PAN, which involves identifying sentence-level style changes in multi-author documents under varying topical constraints. We employ a RoBERTa-based model that captures subtle stylistic differences between consecutive sentences. Our approach achieves F1 scores of 0.823, 0.766, and 0.667 on the easy, medium, and hard datasets, respectively, demonstrating robust performance across increasing levels of difficulty.

Keywords

PAN 2025, Multi-Author Writing Style Analysis, Style Change Detection, Authorship Attribution, RoBERTa, Pre-trained Language Model, Transformers, Fine-Tuning, Stylometric Analysis, CLEF 2025

1. Introduction

The objective of the style change detection task is to locate the places in a multi-author document where authorship shifts. [1] This raises a key question in authorship analysis: is it possible to find stylistic evidence that many authors are present in a text that was authored collaboratively? Solving this problem is particularly relevant in scenarios where reference documents are unavailable, as it enables plagiarism detection purely through internal stylistic cues. Beyond that, style change detection has practical applications in verifying claimed authorship, identifying instances of ghostwriting or gift authorship, and supporting tools for collaborative writing.

Over the years, the PAN shared task on multi-author writing style analysis has evolved significantly. Earlier editions focused on identifying whether a document was authored by one or more individuals (2018), estimating the number of contributing authors (2019) [3], and detecting style changes at the paragraph level (2020–2022) [4] [5]. In 2022, the task was extended to pinpoint style changes at the sentence level [6]. Until then, many of the datasets exhibited high topical variability, which allowed participants to exploit shifts in content as proxies for style variation. Recognizing this, the 2023 and 2024 editions placed greater emphasis on controlling for topic, compelling systems to rely on finer-grained stylistic features rather than content-driven cues. [2]

2. Related Work

Previous work on Multi Author Writing Style analysis spans a wide range of methodologies. Multiple efforts adopt a binary classification framework. The document is divided into text segments that are compared to determine whether they are written by the same author or co-written by two different authors. [9, 10, 11, 13, 14]

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

*Corresponding author.

†These authors contributed equally.

hrohra2004@gmail.com (H. Rohra); nirbhay04@gmail.com (N. Shah); sssonawane@pict.edu (S. Sonawane)

Zlatkova et al. explore ensemble learning through the concept of stacking LightGBM classifiers trained over TF-IDF encoding. They also explore the use of Multi Layered Perceptrons for the same. [9]

Weerasinghe et al. explore the task on small and large datasets, with the logistic regression for the former and neural networks for the latter. The authors use metrics such as ROC, AUC and F1-scores to evaluate their models used in TIRA. [10, 18]

Deibel et al. make use of MLP and Bidirectional LSTM architectures. They use the models in an ensemble setup to capture both sequential and non-linear correlations. [11]

A notable approach is the one used by Shams Alshamasi et al. which diverges from the usual binary classification approach to employ a clustering-based approach. The aim is to create clusters of segments sharing the same author. They use algorithms such as K-Means and DBSCAN. [12]

A famous trend in this task is to use transformer architectures such as BERT and RoBERTa for binary classification. [13, 14, 15] We continue on the trend to utilize a fine-tuned RoBERTa model for binary classification. The paper describes our approach in detail.

3. Dataset

To support the development and testing of style change detection models, three datasets corresponding to the three difficulty levels—*easy*, *medium*, and *hard*—are made available. All datasets contain annotated ground truth reflecting sentence-level style changes, with the exception of the last test segment.

Each data set is divided into three segments: the **training set** (70% of the data) is accompanied by ground truth and is used to train and develop models; the **validation set** (15%) also includes ground truth and serves to tune and evaluate model performance; the **test set** (15%) contains no ground truth and is reserved for the final evaluation of submitted systems. This structured split ensures fair benchmarking across all difficulty levels while supporting robust model development and optimization.

Table 1
Dataset Statistics for Each Difficulty Level

Split	Easy		Medium		Hard	
	Documents	Samples	Documents	Samples	Documents	Samples
Training Set	4200	11065	4200	21914	4200	19014
Validation Set	900	2468	900	4590	900	4132

4. Methodology

This section describes the methodology for implementation of our approach. We approached the problem as a binary classification task, predicting whether a particular sentence pair shares the same authorship or not. Our goal was to leverage the capabilities of pre-trained language models to achieve stable precision and recall results on both seen and unseen data. To achieve this goal it was necessary to prepare training data and conduct fine tuning in a manner to avoid model overfitting. The approach can be explained in 2 phases: 1) Data Preparation and 2) Model fine tuning.

4.1. Data Preparation

To prepare the dataset for fine-tuning, we created sentence pairs from the input documents. Using the provided solution vector, we identified sentence pairs written by the same author (negative class) and by multiple authors (positive class). However, the resulting dataset was overwhelmingly unbalanced, leading to a bias toward same-authored pairs. To address this, we divided our documents into training and validation sets. From each training document, we sourced an equal number of positive and negative sentence pairs to ensure the training data was balanced. For the validation set, we included all possible sentence pairs to ensure the model is evaluated on data that closely resembles realistic inputs.

To prepare sentence pairs for our model we performed tokenization using pretrained tokenizers for the model architecture(RoBERTa in our case). For our truncation strategy we employed the maximum token length to be 256 tokens with the assumption that individual sentences would rarely cross the length restrictions. This was a practical adaptation that allowed us to balance between capturing sufficient contextual information and maintaining computational efficiency.

4.2. Model Fine-Tuning

The pretrained RoBERTa-base [7] model was chosen to be the base model for our approach. RoBERTa has the ability to develop rich linguistic representations for high performance on downstream tasks.

To avoid overfitting and improve generalisation we employed weight decay regularization [8]. Weight decay penalizes large weights in the model effectively adding a L2 Regularization term. This prevented the model from simply memorizing author-specific quirks and focus on comparing stylometric features of sentence pairs.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda \sum_i w_i^2$$

where $\mathcal{L}_{\text{CE}}$ is the standard cross-entropy loss, λ is the weight decay coefficient, and w_i are the trainable weights of the model.

For accelerated training and efficient hardware usage we utilized fp16 for our training that allowed us to train weights on 16 bit floating point arithmetic instead of 32 bit standard precision. We relied on dynamic loss scaling to maintain training stability despite the lower precision.

4.3. Classification Head

The model utilizes a standard classification head to generate our final predictions. It consists of a Dropout Layer attached to the final hidden layers to counter overfitting, a linear layer and an activation layer where we utilized softmax for inference.

Table 2
Architecture of the Default Classification Head

Component	Description
Dropout Layer	Introduced before the classification layer to reduce overfitting by randomly zeroing out elements of the hidden state during training.
Fully Connected Linear Layer	Maps the pooled sentence representation to a 2-dimensional output space corresponding to the binary classes.
Softmax Activation	Converts the output logits into class probabilities for binary classification during inference.

This architecture enables the model to learn the subtle stylistic differences between sentence pairs and make accurate binary predictions on authorship change.

5. Experimentation

This section outlines the experimental setup used for training and evaluation, followed by the results obtained on the official PAN 2025 datasets [16, 17] for the easy, medium, and hard difficulty levels.

5.1. Experimental Setup

We fine-tuned the `roberta-base` model using the Hugging Face `Trainer` API. The model was trained using the settings summarized in Table 3. Training was conducted on a Tesla P100 GPU with mixed precision enabled (`fp16=True`) to optimize GPU memory and computation time.

Table 3
Training Configuration

Parameter	Value
Model	`roberta-base`
Batch Size (Effective)	128 (via gradient accumulation)
Max Token Length	256
Learning Rate	1e-6
Optimizer	AdamW
Weight Decay	0.01
Epochs	5
Precision	Mixed (fp16)
Loss Function	Cross-Entropy with L2 regularization
Evaluation Strategy	Per epoch
Best Model Selection	Based on validation loss

Balanced sentence pairs (equal positive and negative) were used for training, while all sentence pairs were used in validation to mimic real-world data distributions. The model was checkpointed after each epoch, and the best-performing checkpoint was used for final inference.

5.2. Results

We evaluated the model's ability to detect style changes across all three PAN-provided difficulty levels: easy, medium, and hard. The results are presented in Table 4. Evaluation metrics include precision, recall, and F1-score.

Table 4
Comparison of Our Method with Baselines across All Tasks

Approach	Task 1 (Accuracy)	Task 2 (F1)	Task 3 (F1)
Our Method	0.823	0.766	0.667
Baseline Predict 1	0.466	0.343	0.320
Baseline Predict 0	0.112	0.323	0.346

The model performs strongly on the *easy* and *medium* datasets, where topic variation provides subtle clues for authorship change. In the *hard* setting—where topical consistency is enforced—the model still maintains reasonable performance, demonstrating its ability to identify fine-grained stylistic differences without relying on content-based signals.

6. Conclusion

In this paper, we focused on the PAN 2025 task of detecting style change at the sentence level in multi-author writings. Our solution frames the task as a binary classification problem, using the RoBERTa-base model fine-tuned on balanced sentence pairs. We proposed data balancing, regularization, and precision-aware training strategies to enhance the model's robustness and generalizability.

Our model scored F1 of 0.823, 0.766, and 0.667 on the easy, medium, and hard datasets respectively, improving upon naive baselines and providing competitive performance even under stringent topical limitations. These outcomes point to the model's capacity for sensitive stylistic variations without relying substantially upon topic changes. Future directions could leverage more advanced architectures or contrastive learning strategies towards improved performance, particularly in topic-consistent settings.

Acknowledgments

We thank the PAN organizers for providing the datasets and evaluation framework. We also acknowledge the computational resources provided by Pune Institute of Computer Technology.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] E. Zangerle, M. Mayerl, M. Potthast, et al., *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.

[2] J. Bevendorff, X. B. Casals, B. Chulvi, et al., *Overview of PAN 2024: Multi-Author Writing Style Analysis, Multilingual Text Detoxification, Oppositional Thinking Analysis, and Generative AI Authorship Verification*, in: L. Goeuriot et al. (Eds.), *Proceedings of the Fifteenth International Conference of the CLEF Association (CLEF 2024)*, Lecture Notes in Computer Science, Springer, 2024.

[3] E. Zangerle, M. Tschuggnall, G. Specht, et al., *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.

[4] E. Zangerle, M. Mayerl, G. Specht, et al., *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.

[5] E. Zangerle, M. Mayerl, M. Potthast, et al., *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.

[6] E. Zangerle, M. Mayerl, M. Potthast, et al., *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.

[7] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, *RoBERTa: A Robustly Optimized BERT Pretraining Approach*, arXiv preprint arXiv:1907.11692, 2019.

[8] A. Kosson, B. Messmer, and M. Jaggi, *Rotational Equilibrium: How Weight Decay Balances Learning Across Neural Networks*, arXiv preprint arXiv:2305.17212, 2023.

[9] D. Zlatkova, D. Kopev, K. Mitov and A. Atana, *An ensemble rich multi-aspect approach for robust style change detection*, *PAN at CLEF*, 2018.

[10] J. Weerasinghe and R. Greenstadt, *Feature Vector Difference based Neural Network and Logistic Regression Models for Authorship Verification*, *PAN at CLEF*, 2020.

[11] R. Deibel and D. Löfflad, *Style Change Detection on Real-World Data using an LSTM-powered Attribution Algorithm*, *PAN at CLEF*, 2021.

[12] S. Alshamasi and M. Menai, *Ensemble-Based Clustering for Writing Style Change Detection in Multi-Authored Textual Documents*, *PAN at CLEF*, 2022.

[13] Z. Chen, Y. Han, and Y. Yi, *Team Chen at PAN: Integrating R-Drop and Pre-trained Language Model for Multi-author Writing Style Analysis*, *PAN at CLEF*, 2024.

[14] A. A. Khan, M. Rai, K. A. Khan, S. J. Shah, F. Alvi, and A. Samad, *Team Gladiators at PAN: Improving Author Identification: A Comparative Analysis of Pre-Trained Transformers for Multi-Author Classification*, *PAN at CLEF*, 2024.

[15] A. Wegmann, M. Schraagen, and D. Nguyen, *Same Author or Just Same Topic? Towards Content-Independent Style Representations*, *Proceedings of the 7th Workshop on Representation Learning for NLP, ACL*, Dublin, Ireland, 2022, pp. 249–268.

[16] 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, and E. Zangerle, *Overview of PAN 2025: Voight-Kampff Generative AI Detection, Multilingual Text Detoxification, Multi-Author Writing Style Analysis, and Generative Plagiarism Detection, CLEF 2025, Lecture Notes in Computer Science, Springer*, Madrid, Spain, 2025.

[17] E. Zangerle, M. Mayerl, M. Potthast, and B. Stein, *Overview of the Multi-Author Writing Style Analysis Task at PAN 2025, Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings*, Madrid, Spain, 2025.

[18] M. Fröbe, M. Wiegmann, N. Kolyada, B. Grahm, T. Elstner, F. Loebe, M. Hagen, B. Stein, and M. Potthast, *Continuous Integration for Reproducible Shared Tasks with TIRA.io*, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), *Advances in Information Retrieval - 45th European Conference on Information Retrieval (ECIR 2023), Dublin, Ireland, April 2–6, 2023, Proceedings, Part III*, Lecture Notes in Computer Science, vol. 13982, Springer, 2023, pp. 236–241.