CLEF 2025 PAN: Multi-Author Writing Style Analysis

Ali Muhammad Asad[1], Musaib[2], Syed Muhammad Areeb Kazmi[3] and Sarim Tahir[4]

[1]*Department of Computer Science, Habib University*
[2]*Department of Computer Science, Habib University*
[3]*Department of Computer Science, Habib University*
[4]*Department of Computer Science, Habib University*

Abstract
Multi-Author Writing Style Analysis is a crucial task in computational linguistics and authorship attribution, aimed at identifying points of transition within a document. This task has significant applications in forensic linguistics, plagiarism detection, and content verification. Over the years, various approaches have been explored, ranging from traditional stylometric techniques to advanced deep learning models. Early research relied on lexical and syntactic features to detect style changes, but these methods faced limitations in handling nuanced transitions. Recent advancements, particularly the integration of transformer-based models, have significantly improved the accuracy of style change detection. This paper aims to provide an overview of the tasks, the datasets, the evaluation metrics, and the baseline models for the Multi-Author Writing Style Analysis Lab of the PAN track at CLEF 2025.

1. Introduction

Writing style analysis is a fundamental problem in computational linguistics, with applications in authorship attribution, forensic text analysis, and collaborative content verification. A challenge in this domain is detecting when authorship changes within a multi-author document, a task known as style change detection. This problem has been the focus of multiple iterations of the PAN shared tasks, which have provided benchmark datasets and evaluation frameworks to advance the field.

Early approaches to style change detection leveraged handcrafted features such as n-grams, part-of-speech tag frequencies, and sentence structure analysis. These methods, while effective for basic segmentation tasks, struggled with complex, subtle changes in authorial style. The introduction of deep learning and pre-trained transformer models revolutionized the field, allowing for more nuanced analysis of writing styles. Models like BERT and DeBERTa have demonstrated superior performance in detecting stylistic shifts at both the paragraph and sentence levels.

This paper aims to explore existing methods and challenges in multi-author writing style analysis, reviewing state-of-the-art approaches such as deep learning models, transfer learning, and ensemble techniques. We will discuss the datasets, evaluation metrics, and baseline models used in the previous iterations of Multi-Author Writing Style Analysis Lab of the PAN track, ultimately attempting to improve the performance of style change detection.

2. Literature Review

Traditional approaches to multi-author writing style analysis mainly rely on lexical and syntactic feature based methods such as TF-IDF vectorization, character n-grams, and POS tag frequencies. These methods have demonstrated moderate performance in detecting style change at the paragraph level, but struggle with sentence-level changes due to limited contextual understanding. For instance, the PAN 2021 task utilized lexical similarity measures to detect authorship changes, achieving an F1 score of 0.73 on multi-author documents, indicating that feature-based methods were more effective for

aa07190@st.habib.edu.pk (A. M. Asad); mu07694@st.habib.edu.pk (Musaib); mk07202@st.habib.edu.pk (S. M. A. Kazmi); st07112@st.habib.edu.pk (S. Tahir)

document-wide analysis but struggled with fine-grained author attribution [1]. PAN 2022 introduced structural features such as indentation and average sentence length, further refining style change detection by incorporating discourse-level analysis [2]. However, these methods were constrained by their reliance on handcrafted features and were susceptible to topic influence.

PAN 2023 and 2024 introduced more powerful techniques such as transformer based methods. The top-performing models in PAN 2023 leveraged pre-trained transformers like DeBERTaV3 and BERT for paragraph-level classification, achieving state-of-the-art F1 scores of 0.83 for the hardest dataset variant [3]. These models incorporated contrastive learning and fine-tuning on task-specific datasets, allowing them to distinguish between subtle stylistic variations more effectively. However, their performance degraded when topic variations were minimized, highlighting the challenge of isolating style from content.

Another approach integrated contrastive learning with DeBERTaV3 and knowledge distillation for efficiency [3]. This method allowed models to refine their representations of stylistic shifts while maintaining computational feasibility. Additionally, multi-stage fine-tuning techniques were introduced in PAN 2024, further improving performance by integrating both syntactic and semantic embeddings for style differentiation [4]. Hybrid approaches were also explored, combining traditional and deep learning methods. One study used a stacking ensemble of lexical and neural models, achieving an F1 score of 0.82 on medium-difficulty datasets [3]. Another team experimented with sentence-level style change detection using hierarchical clustering and transformer embeddings, demonstrating the potential for finer granularity in authorship attribution [4].

Performance evaluation has been standardized using the macro-averaged F1-score, with separate rankings for easy, medium, and hard datasets. PAN 2023 results highlighted the superiority of transformer-based models over traditional techniques, with the best-performing systems leveraging multi-stage fine-tuning, semantic similarity post-processing, and ensemble classifiers [4]. However, performance variations across dataset difficulty levels suggest that improvements in data augmentation and domain adaptation are needed. Another paper in PAN 2024 explored bagging techniques, feature engineering, BERT-based classifiers and ensemble techniques combining different architectures such as BERT, RoBERTa, Electra and Llama2, using DetectGPT (Mistral-7B Falcon-7B) as their baseline models. They were able to achieve an F1-score of 0.924 as their best results on fine-tuned Mistral and Llama2 models [5].

3. Methodology

3.1. Dataset

The dataset was provided by PAN, based on user posts from various subreddits of the Reddit platform. It is divided into three levels: easy, medium and hard, where each level is split into three parts:

- *training set:* Contains 70% of the whole dataset and includes ground truth data. This data would be used to develop and train the models.
- *validation set:* Contains 15% of the whole dataset and includes ground truth data. This data would be used to evaluate and optimize the models.
- *test set:* Contains 15% of the whole dataset and does not include ground truth data. This data would be used to evaluate the models.

Input Format

For each problem instance X (i.e., each input document), two files are provided:

1. *problem-X.txt* which contains the actual text in the form of sentences of varying lengths.
2. *truth-problem-X.json* which contains the ground truth, i.e., the correct solution in JSON format.

A sample json file looks as so:

```
{"authors": 2,
"changes": [0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0]}
```


where the key ‘changes’ is an array of consecutive sentences within each document (0 when there is no change, and 1 when there is a change).

3.2. Initial Approach

The initial approach, which is also our baseline, was to use the Bag of Words (BoW) method for feature extraction from the sentences, coupled with a simple logistic regression model for classification. The Bag of Words method is a simple and effective way to represent text data. It involves creating a vocabulary of unique words and then representing each sentence as a vector of word counts. This became our baseline approach, as it is simple and easy to implement, and was backed by previous implementations [5].

Once our baseline was set, we modified the bag of words approach to combine n-grams and syntactic/lexical features (sentence length, POS tag frequencies, etc), along with calss weighting to handle the class imbalance between instances of 0s and 1s, and finally used a vectorized tuning with ngram_range of (1, 2) to include both unigrams and bigrams in the feature set, in order to improve the model’s performance by capturing more context and semantics from the text data. This showed an improvement in the model’s performance over the baseline.

4. Results

This section presents the results of our experiments. The table below shows the performance of the models on the validation set. The models were evaluated based on the F1 score, which is the harmonic mean of precision and recall. The F1 score is a good metric for imbalanced datasets, as it considers both false positives and false negatives. The F1 score is presented for each level, easy, medium and hard.

Level	**F1 (Easy)**	**F1 (Medium)**	**F1 (Hard)**
Baseline BoW	0.65	0.65	0.57
Improved BoW	0.79	0.67	0.61

Table 1
F1 scores of the models on the validation set

The results show an improvement in the F1 score for the improved BoW model over the baseline BoW model on all levels.

References

[1] E. Zangerle, M. Mayerl, M. Potthast, B. Stein, Overview of the style change detection task at pan 2021, CLEF 2021 - Conference and Labs of the Evaluation Forum, September 21-24, 2021, Bucharest, Romania (2021).

[2] E. Zangerle, M. Mayerl, M. Potthast, B. Stein, Overview of the style change detection task at pan 2022, CLEF 2022 - Conference and Labs of the Evaluation Forum, September 5-8, 2022, Bologna, Italy (2022).

[3] E. Zangerle, M. Mayerl, M. Potthast, B. Stein, Overview of the multi-author writing style analysis task at pan 2023, CLEF 2023 - Conference and Labs of the Evaluation Forum, September 18-21, 2023, Thessaloniki, Greece (2023).

[4] E. Zangerle, M. Mayerl, M. Potthast, B. Stein, Overview of the multi-author writing style analysis task at pan 2024, CLEF 2024: Conference and Labs of the Evaluation Forum, September 09-12, 2024, Grenoble, France (2025).

[5] A. A. Ayele, N. Babakov, J. Bevendorff, X. B. Casals, B. Chulvi, D. Dementieva, A. Elnagar, D. Freitag, M. Fröbe, D. Korencic, M. Mayerl, D. Moskovskiy, A. Mukherjee, A. Panchenko, M. Potthast, F. Rangel, N. Rizwan, P. Rosso, F. Schneider, A. Smirnova, E. Stamatatos, E. Stakovskii, B. Stein, M. Taulé, D. Ustalov, X. Wang, M. Wiegmann, S. M. Yimam, E. Zangerle, Overview of pan 2024: Multi-author writing style analysis, multilingual text detoxification, oppositional thinking analysis, and generative ai authorship verification condensed lab overview, CLEF 2024: Conference and Labs of the Evaluation Forum, September 09-12, 2024, Grenoble, France (2024).