

# Enhancing Writing Style Change Detection using Transformer-based Models and Data Augmentation

Notebook for PAN at CLEF 2023

Ahmad Hashemi<sup>1</sup>, Wei Shi<sup>1</sup>

<sup>1</sup>*School of Information Technology, Carleton University, Ottawa, Canada*

## Abstract

This paper presents our participation in the Style Change Detection task for PAN at CLEF 2023. The primary goal of this task is to identify alterations in writing style at the paragraph level within a provided document that has been authored by multiple writers. The task comprises three sub-tasks that differ in difficulty levels, primarily based on the diversity of topics addressed within the paragraphs. To address these sub-tasks, we investigate the effectiveness of fine-tuning different pre-trained transformer-based models, with a particular emphasis on RoBERTa. Additionally, we employ data augmentation techniques to enhance the performance of our models. Furthermore, we incorporate ensemble modeling to further improve the accuracy and robustness of our style change detection system. In the competition, our provided models achieved the first rank in terms of F1 score for two of the sub-tasks, and secured the second position for the remaining sub-task.

## Keywords

Authorship Attribution, Plagiarism Detection, Ensemble Learning, Transformers

## 1. Introduction

Multi-Author writing style analysis is an interesting area of study that focuses on analyzing documents that have been written by multiple authors. It involves a range of tasks, such as determining whether a document is the product of a single author or multiple authors [1], as well as investigating the occurrence and positioning of style changes in multi-authored documents [2]. The primary motivation behind Multi-Author writing style analysis is that it enables the identification of positions where authors switch within a text, allowing for the detection of plagiarism and the verification of claimed authorship, even in the absence of comparison texts. Style change detection also assists in uncovering gift authorships and can contribute to the development of innovative technologies for writing support [3].

The style change detection task introduced by PAN [4] this year focuses on detecting writing style changes at the paragraph level in a given text document. The objective is to detect style changes between consecutive paragraphs, assessing whether there was a transition in writing style. The task provides datasets of three difficulty levels: easy (subtask 1), medium (subtask 2), and hard (subtask 3). In the easy level, paragraphs cover a range of topics, allowing approaches to

---


*CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece*

✉ [ahmadhashemi@cmail.carleton.ca](mailto:ahmadhashemi@cmail.carleton.ca) (A. Hashemi); [wei.shi@carleton.ca](mailto:wei.shi@carleton.ca) (W. Shi)

🆔 0000-0003-2853-4963 (A. Hashemi); 0000-0002-3071-8350 (W. Shi)



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

utilize topic information for detecting authorship changes. The medium level features a smaller topical variety, requiring the approach to focus more on style to effectively solve the detection task. The hard level consists of paragraphs on the same topic throughout the document [5].

## 2. Related work

With the advancements in natural language processing techniques, many researchers have directed their attention toward various tasks within the realm of digital text forensics. These tasks cover a wide range, including the detection of fake news [6], spam [7], and hate speech [8], as well as author profiling [9], authorship attribution [10], and style change detection [11]. Earlier studies primarily employed feature-based approaches, involving the extraction of stylistic features followed by applying traditional machine learning or deep learning algorithms. However, the emergence of pre-trained transformer-based models, with their remarkable capabilities, has led to a shift in focus for many recent studies. These studies now mostly center around fine-tuning pre-trained models to customize them for their respective tasks.

In the domain of style change detection, early attempts primarily revolved around the extraction of stylometric features. For instance, Eissen and Stein [12] employed word frequency classes to differentiate between distinct writing styles in order to investigate intrinsic plagiarism detection. Bensalem et al. [13] utilized n-grams to identify authorial style changes, while Gianella [14] applied Bayesian modeling techniques to segment a document based on authorship. Another approach [15], involved the use of neural networks in conjunction with various stylometric features.

More recent approaches in style change detection have mostly leveraged pre-trained models, although some still incorporate stylometric features. For example, Iyer and Vosoughi [16] employed Google AI’s BERT pre-trained bidirectional models to tokenize and generate sentence embeddings, which were then utilized to train a random forest classifier for the PAN’s SCD task. For the PAN SCD task of 2021, Singh et al. [17] extracted stylometric features from each paragraph and used the absolute differences between the feature vectors to train a Logistic Regression classifier to determine if two paragraphs were written by the same author. Lin et al. [18] fine-tuned transformer models such as BERT, RoBERTa, and ALBERT, along with their classifiers, to measure the similarity between paragraphs or sentences for authorship analysis in the most recent PAN style change detection task in 2022.

In our work, we investigate the effectiveness of leveraging three pre-trained transformer-based models in detecting style changes between consecutive paragraphs, where both the author and topic change, as well as cases where only the author changes while the topic remains the same. Additionally, we apply data augmentation techniques and employ ensemble modeling to enhance the performance of our approach.

## 3. Dataset

For each of the subtasks, a separate dataset has been provided. These datasets comprise multiple documents, with each document containing some paragraphs. For every document, a corresponding ground truth file is available, providing two pieces of information: 1) the number

**Table 1**

Datasets statistics. "Training 1," "Training 2," and "Training 3" correspond to the tasks with easy, medium, and hard difficulty levels, respectively. Similarly, "Validation 1," "Validation 2," and "Validation 3" represent the validation sets for each respective task.

Dataset	documents	Avg. paragraphs per document	samples	positive	negative
Training 1	4200	4.07	12,904	11,347	1,557
Training 2	4200	7.71	28,216	13,215	15,001
Training 3	4200	5.55	19,113	9,021	10,092
Validation 1	900	4.14	2,828	2,451	377
Validation 2	900	8.82	7,042	3,029	4,013
Validation 3	900	5.56	4,112	1,953	2,159

of authors associated with the document, and 2) the identification of consecutive paragraphs where a style change has occurred, indicating a transition in authorship. Each subtask's dataset has been divided into a training set and a validation set.

For our experimental setup, we treated every pair of consecutive paragraphs as a sample and concatenated them. Each sample was assigned a label indicating whether a style change occurred between the two paragraphs (labeled as 1) or if they were written by the same author (labeled as 0). Further details and statistics about the datasets in our experimental setup can be found in Table 1.

## 4. Methodology

### 4.1. Data preparation

To create the samples, we begin by concatenating two consecutive paragraphs within each document using a separator token. Next, we assign the associated binary label indicating whether a style change occurs between the two paragraphs. This allows us to transform the task into a binary classification problem. To prepare the samples for fine-tuning the pre-trained transformer-based models (specifically BERT, RoBERTa, and ELECTRA), we employ the corresponding tokenizer associated with each model. However, it is important to note that these models have limitations regarding the maximum input sequence size, typically set at 512 tokens. To address this limitation, we analyze the datasets and find that in each dataset only a few samples exceed the maximum token limit. Therefore, we opt to truncate the longer samples. To ensure equal attention is given to each paragraph in the sample, we adopt a truncation strategy that involves removing tokens from both ends of the sequence.

### 4.2. Pre-trained Transformer Models

Pre-trained transformer models are powerful language models that are trained on massive amounts of text data. They learn to understand the structure and patterns of language, enabling them to generate high-quality text and perform various natural language processing (NLP) tasks. Fine-tuning the pre-trained models allows us to leverage their language understanding

capabilities and transfer the knowledge they have acquired from their extensive pre-training to our specific task. In our study, we employed three popular pre-trained transformer models, namely BERT, RoBERTa, and ELECTRA.

BERT (Bidirectional Encoder Representations from Transformers) is a revolutionary model that introduced the concept of bidirectional context to capture the dependencies between words. It utilizes a transformer architecture and pre-training tasks such as masked language modeling to learn contextualized representations of words [19]. RoBERTa, an extension of BERT, further enhanced the pre-training process by utilizing additional training data and applying dynamic masking strategies. This enabled RoBERTa to achieve even better performance across a range of NLP tasks [20]. ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately) introduced a novel pre-training method called "discriminative masked language modeling." It improves efficiency by generating synthetic training data and training the model to differentiate between original and replaced tokens [21].

To adapt these pre-trained models for our specific task of style change detection, we added a binary classification layer on top of each model. This allowed the models to learn and classify whether a style change occurred between the consecutive paragraphs in each sample. For each subtask, we fine-tuned the models using the associated dataset to consider the unique characteristics of each subtask.

### 4.3. Data augmentation

To explore the potential impact of generating more samples while disregarding the consecutiveness and order of the paragraphs, we conduct an investigation. We transform the task into determining whether the same author wrote two concatenated paragraphs (not necessarily consecutive or in order). This allows us to explore different approaches for data augmentation. In our first approach, we swap the order of the two paragraphs in each sample, creating a new sample with the same label to double the number of samples. Additionally, we leverage the metadata provided in the datasets to identify paragraphs written by the same authors that are not necessarily consecutive. Using the metadata, which includes the number of authors in each document, we employ the Algorithm 1 for each document to retrieve such pairs of paragraphs and augment the data.

As Algorithm 1 demonstrates, we incorporate additional non-consecutive pairs of paragraphs into our sample set and assign them labels based on the inferred relationships. For example, if there are three consecutive paragraphs without a style change, we can infer that the first and third paragraphs are written by the same author. Similarly, if there are style changes between the first and second paragraphs and between the second and third paragraphs, we can deduce that the authors of the first and third paragraphs are different, given that the number of authors in the document exceeds the number of style changes by one.

After data augmentation, we utilize our extended sample sets for each subtask to fine-tune the pre-trained model following the same process explained earlier. It is important to note that this augmentation introduces a change in the nature of the train set, deviating from the consecutive and ordered structure that exists in the test data. However, the advantage of having a larger dataset can improve the model's ability to capture patterns across paragraphs written by the same author.

---

**Algorithm 1:** Pseudocode for data augmentation based on non-consecutive paragraph pairs

---

```
if style_changes.count(1) = (authors_count - 1) then
  for  $i$  in range(len(paragraphs) - 1) do
     $j \leftarrow i + 1$ ; // set the next paragraph index
    while  $j < \text{len}(\text{paragraphs})$  and style_changes[ $j - 1$ ] = 0 do
      ; // while same author
      if  $j > i + 1$  then
        samples.append(paragraphs[ $i$ ] + separator + paragraphs[ $j$ ])
        labels.append(0); // same author
      end
       $j \leftarrow j + 1$ ; // move to the next paragraph
    end
    while  $j < \text{len}(\text{paragraphs})$  do
      if  $j > i + 1$  then
        samples.append(paragraphs[ $i$ ] + separator + paragraphs[ $j$ ])
        labels.append(1); // style change
      end
       $j \leftarrow j + 1$ ; // move to the next paragraph
    end
  end
end
```

---

#### 4.4. Generalization and Ensemble modeling

Given the similarity in the nature of datasets across different subtasks, particularly in terms of the presence of style changes, we explore the potential benefits of leveraging datasets from other subtasks to enhance the model performance for a specific subtask. For example, Task 2 and Task 3 share similarities as they both involve style changes occurring while the topic remains consistent. Similarly, Task 2 and Task 1 exhibit similarities as they both encompass style changes alongside topic transitions. We believe that the datasets from Task 1 and Task 3 can provide valuable insights to Task 2, as they encompass scenarios that align with this task. Accordingly, our investigation involves not only fine-tuning the model using the provided dataset specific to the subtask but also incorporating additional samples from other subtasks' datasets to assess the impact of having a generalized model.

To maximize the potential benefits offered by the various approaches mentioned, we employ an ensemble strategy based on the majority voting approach. For each subtask, we develop three models, all based on fine-tuning the RoBERTa pre-trained model, which we will discuss in subsequent sections as the most effective among the investigated pre-trained models for all subtasks. The first model utilizes the initial samples exclusively from the task-specific dataset for fine-tuning. The second model incorporates augmented samples derived from the task-specific dataset, thereby expanding the training data. Lastly, the third model combines all the original

samples from the other datasets in addition to the task-specific dataset to provide a generalized model. By ensembling these models, we aim to leverage the strengths of each approach and enhance the overall performance of our style change detection system.

## 5. Experiments

### 5.1. Experimental settings

We downloaded the large versions of pre-trained BERT, RoBERTa, and ELECTRA models from HuggingFace [22]. The implementation and fine-tuning of these models were conducted on a server equipped with an NVIDIA A100 GPU. To optimize the performance of our models, we selected hyperparameter values as follows: We set the maximum sequence length to 512, the learning rate to 0.00001, the batch size to 16, and the number of epochs to 10.

To assess the effectiveness of the models for each subtask, we evaluate their performance by computing the F1 score on the provided evaluation set. The F1 score is calculated based on the predictions made by the models for detecting style changes between consecutive paragraph pairs within the evaluation set of each subtask. After conducting our experiments and obtaining results on the evaluation sets, we select the best-performing model for each subtask and run the model on an unseen test set using the TIRA platform [23].

### 5.2. Results

The results of our experiments are presented in this section. Firstly, Table 2 displays the performance comparison of fine-tuning pre-trained models on the original task-specific datasets for each subtask. The findings indicate that RoBERTa consistently outperforms both BERT and ELECTRA across all subtasks, establishing its superiority in the task of style change detection. As a result, we have selected RoBERTa as the preferred approach for our style change detection system.

Moving on, Table 3 presents the performance evaluation results of our RoBERTa-based experiments on the evaluation set. It highlights the F1 scores for each subtask and approach. The findings reveal that fine-tuning RoBERTa on the original task-specific dataset yields the highest F1 scores among all the provided approaches for subtasks 1 and 3. As can be seen, the performance of the generalized model trained on all the datasets drops for subtask 1 and subtask 3, which aligns with our expectations as these subtasks possess exclusive characteristics that may not benefit from additional data from the other subtasks. On the other hand, subtask 2 exhibits similarities to both subtask 1 (style changes along with topic changes) and subtask 3 (style changes without topic change), which explains why the performance drop for the generalized model is less significant in subtask 2.

Furthermore, incorporating augmented data for subtask 3 leads to a notable performance drop, suggesting that considering paragraph order and consecutiveness is not negligible for detecting style changes when the topic is consistent. However, the utilization of augmented data produces competitive performance compared to only using the original samples for subtasks 1 and 2. Notably, the ensemble model for subtask 2 achieves the best F1 score, indicating that the

**Table 2**

F1 score obtained by fine-tuning each pre-trained model for each of the subtasks on the associated validation set. The best result for each dataset is given in bold.

Pre-trained Model	Subtask 1	Subtask 2	Subtask 3
BERT	0.9823	0.7451	0.7048
ELECTRA	0.9882	0.8024	0.7797
RoBERTa	<b>0.9957</b>	<b>0.8106</b>	<b>0.8140</b>

**Table 3**

F1 score obtained by each RoBERTa-based approach for each of the subtasks on the associated validation set. The best result for each dataset is given in bold.

Approach	Subtask 1	Subtask 2	Subtask 3
Basic	<b>0.9957</b>	0.8106	<b>0.8140</b>
Augmented data	0.9832	0.8053	0.7266
Generalized (all datasets)	0.9661	0.8042	0.7449
Ensemble	0.9907	<b>0.8221</b>	0.7906

**Table 4**

Final performance results on the unseen test sets.

Task	Reported F1 Score
Subtask 1	0.984
Subtask 2	0.843
Subtask 3	0.812

individual models capture complementary patterns that contribute to the overall performance improvement.

Based on the validation set results, we select the ensemble approach for subtask 2 and the basic RoBERTa models for subtasks 1 and 3 to perform on the final test sets. The final results of our selected models on the unseen test sets are presented in Table 4. According to the provided results reported by the competition organizers, our provided models for subtasks 1 and 2 outperformed all other participants’ models, securing the first position. For subtask 3, we achieved the second position.

## 6. Conclusion

In conclusion, our investigation aimed to enhance style change detection in textual documents through various techniques and approaches. We explored data augmentation strategies to generate non-consecutive paragraph pairs, allowing the model to learn patterns beyond sequential and ordered paragraphs. We found that fine-tuned RoBERTa models outperformed BERT and ELECTRA in all the task difficulty levels, demonstrating their effectiveness in capturing style changes. The results also highlighted the importance of considering the unique nature of

each task, as incorporating additional data from unrelated tasks did not necessarily improve performance.

Furthermore, the ensemble approach proved to be valuable in capturing complementary patterns, particularly for subtask 2, where the nature of the subtask contained similarities to both the other subtasks. This ensemble model outperformed individual models, emphasizing the benefits of combining multiple perspectives. We believe these insights contribute to advancements in authorship attribution and plagiarism detection applications.

## References

- [1] 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.
- [2] E. Zangerle, M. Mayerl, G. Specht, M. Potthast, B. Stein, Overview of the style change detection task at pan 2020., CLEF (Working Notes) 93 (2020).
- [3] Style Change Detection Task at CLEF 2023, <https://pan.webis.de/clef23/pan23-web/style-change-detection.html>, 2023. Accessed: July 2, 2023.
- [4] J. Bevendorff, I. Borrego-Obrador, M. Chinea-Ríos, M. Franco-Salvador, M. Fröbe, A. Heini, K. Kredens, M. Mayerl, P. Pęzik, M. Potthast, F. Rangel, P. Rosso, E. Stamatatos, B. Stein, M. Wiegmann, M. Wolska, , E. Zangerle, Overview of PAN 2023: Authorship Verification, Multi-Author Writing Style Analysis, Profiling Cryptocurrency Influencers, and Trigger Detection, in: A. Arampatzis, E. Kanoulas, T. Tsikrika, A. G. Stefanos Vrochidis, D. Li, M. Aliannejadi, M. Vlachos, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2023), Lecture Notes in Computer Science, Springer, 2023.
- [5] 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 CLEF 2023 - Conference and Labs of the Evaluation Forum, CEUR-WS, 2023.
- [6] X. Zhou, R. Zafarani, A survey of fake news: Fundamental theories, detection methods, and opportunities, ACM Computing Surveys (CSUR) 53 (2020) 1–40.
- [7] S. Rao, A. K. Verma, T. Bhatia, A review on social spam detection: Challenges, open issues, and future directions, Expert Systems with Applications 186 (2021) 115742.
- [8] W. Yin, A. Zubiaga, Towards generalisable hate speech detection: a review on obstacles and solutions, PeerJ Computer Science 7 (2021) e598.
- [9] Y. HaCohen-Kerner, Survey on profiling age and gender of text authors, Expert Systems with Applications (2022) 117140.
- [10] W. Zheng, M. Jin, A review on authorship attribution in text mining, Wiley Interdisciplinary Reviews: Computational Statistics 15 (2023) e1584.
- [11] J. Bevendorff, B. Chulvi, E. Fersini, A. Heini, M. Kestemont, K. Kredens, M. Mayerl,



- R. Ortega-Bueno, P. Pezik, M. Potthast, et al., Overview of pan 2022: Authorship verification, profiling irony and stereotype spreaders, and style change detection, in: *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 13th International Conference of the CLEF Association, CLEF 2022, Bologna, Italy, September 5–8, 2022, Proceedings*, Springer, 2022, pp. 382–394.
- [12] S. M. z. Eissen, B. Stein, Intrinsic plagiarism detection, in: *Advances in Information Retrieval: 28th European Conference on IR Research, ECIR 2006, London, UK, April 10-12, 2006. Proceedings 28*, Springer, 2006, pp. 565–569.
- [13] I. Bensalem, P. Rosso, S. Chikhi, Intrinsic plagiarism detection using n-gram classes, in: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1459–1464.
- [14] C. Giannella, An improved algorithm for unsupervised decomposition of a multi-author document, *Journal of the Association for Information Science and Technology* 67 (2016) 400–411.
- [15] N. Graham, G. Hirst, B. Marthi, Segmenting documents by stylistic character, *Natural Language Engineering* 11 (2005) 397–415.
- [16] A. Iyer, S. Vosoughi, Style change detection using bert., *CLEF (Working Notes)* 93 (2020) 106.
- [17] R. Singh, J. Weerasinghe, R. Greenstadt, Writing style change detection on multi-author documents., in: *CLEF (Working Notes)*, 2021, pp. 2137–2145.
- [18] T.-M. Lin, C.-Y. Chen, Y.-W. Tzeng, L.-H. Lee, Ensemble pre-trained transformer models for writing style change detection, *CLEF*, 2022.
- [19] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, *arXiv preprint arXiv:1810.04805* (2018).
- [20] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, *arXiv preprint arXiv:1907.11692* (2019).
- [21] K. Clark, M.-T. Luong, Q. V. Le, C. D. Manning, Electra: Pre-training text encoders as discriminators rather than generators, *arXiv preprint arXiv:2003.10555* (2020).
- [22] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, et al., Huggingface’s transformers: State-of-the-art natural language processing, *arXiv preprint arXiv:1910.03771* (2019).
- [23] M. Fröbe, M. Wiegmann, N. Kolyada, B. Grahm, T. Elstner, F. Loebe, M. Hagen, B. Stein, M. Potthast, Continuous Integration for Reproducible Shared Tasks with TIRA.io, in: J. Kamps, L. Goeriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), *Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023)*, Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2023, pp. 236–241.