Noname manuscript No.
(will be inserted by the editor)

Stylometry Analysis of Multi-authored Documents for Authorship and Author Style Change Detection

Muhammad Tayyab Zamir ·
Muhammad Asif Ayub · **Asma Gul** ·
Nasir Ahmad · **Kashif Ahmad**

the date of receipt and acceptance should be inserted later

Abstract In recent years, the increasing use of Artificial Intelligence (AI)–based text generation tools has posed new challenges in document provenance, authentication, and authorship detection. However, advancements in stylometry have provided opportunities for automatic authorship and author change detection in multi-authored documents using style analysis techniques. Style analysis can serve as a primary step towards document provenance and authentication through authorship detection. This paper investigates three key tasks of style analysis: (i) classification of single and multi-authored documents, (ii) single change detection, which involves identifying the point where the author switches, and (iii) multiple author-switching detection in multi-authored documents. We formulate all three tasks as classification problems and propose a merit-based fusion framework that integrates several state-of-the-art natural language processing (NLP) algorithms and weight optimization techniques. We also explore the potential of special characters, which are typically removed during pre-processing in NLP applications, on the performance of the proposed methods for these tasks by conducting extensive experiments on both cleaned and raw datasets. Experimental results demonstrate significant improvements over existing solutions for all three tasks on a benchmark dataset.

Muhammad Tayyab Zamir
Abasyn University Islamabad Campus, Pakistan
E-mail: tayyab.zamir999@gmail.com

Muhammad Asif Ayub
Department of Computer Systems Engineering, University of Engineering and Technology, Peshawar, Pakistan.
E-mail: asifayub836@gmail.com

Asma Gul
NUST Business School, National University of Science and Technology, Islamabad, Pakistan.
E-mail: gulasma24@gmail.com

Nasir Ahmad
University of Western Australia, Australia
E-mail: n.ahmad@uetpeshawar.edu.pk

Kashif Ahmad
Department of Computer Science, Munster Technological University Cork, Ireland
E-mail: kashif.ahmad@mtu.ie

Keywords Author Change Detection, Stylometry, Authorship Detection, Author Attribution, Text Classification, LLMs

1 Introduction

In the modern world, the automatic content-generating tools especially the large language models (LLMs) and deep generative AI have made it easier to generate large volumes of grammatically correct and consistent content faster. However, these tools have significantly increased the concerns over the reliability and authenticity of the content in different application domains [19,3]. For instance, in the case of trending news, misinformation or hallucination by such tools may have a significant impact on the policies, trust in governments, and chaos, which may compromise public safety. Similarly, in the education sector, these tools may result in an increase in academic dishonesty in several ways [9].

In the case of textual content, one of the key concerns is authorship identification, which mainly involves assessing the authenticity of a document by identifying its original author(s) [13]. Automatic authorship identification of text documents could prove useful in several application domains such as education (e.g., evaluation and assessment), journalism (e.g., fake news detection and disinformation), law enforcement, and content moderation. Style analysis, which involves exploring and extracting specific patterns in text, is considered a basic step toward developing automatic authorship identification tools. Style analysis enables several interesting tasks and applications allowing the identification of intrinsic features in multi-authored documents. Style change detection (SCD) is one such application of style analysis that could help address growing concerns over document provenance and authentication in various domains [5,6].

The existing literature contains several studies that explore different aspects of SCD, which involve tasks such as the classification of single and multi-authored documents, as well as the identification of paragraphs written by the same or different authors (i.e., the locations where the author switches in a multi-authored document). In this work, our focus is on three key tasks of SCD, namely: (i) classification of single and multi-authored documents, (ii) basic style change (i.e., identification of single author-switching in a document), and (iii) real-world style change (i.e., identification of multiple author-switching in a document). These tasks were also part of a benchmark and competition known as PAN-2021[1].

Several solutions have been proposed as part of the competition. However, there are still key aspects of the tasks that have not been explored yet. While the existing solutions mostly rely on individual NLP models, we believe that combining these models in a merit-based fusion scheme could significantly improve performance. Moreover, special characters such as punctuation keys, question marks, contractions, and stop and short words could play a vital role in style analysis given that their use varies between authors. Although these special characters are typically removed during pre-processing in NLP applications, it would be interesting to explore their potential in SCD given the nature of the application.

To address these open issues, we propose a fusion-based text classification framework by employing multiple transformers and weight selection/optimization methods. We also analyze the impact of special characters on the three SCD tasks by conducting experiments on both clean and raw datasets. Note that this work is an extension of our previous work [27], which tackled only the first task (i.e., classification of single and multi-authored documents).

The main contributions of this work are summarized as follows.

[1] https://pan.webis.de/index.html

- We propose a novel merit-based late fusion framework that tackles the aforementioned three different tasks of stylometry/style change detection within the text.
- We evaluate the performance of various weight selection/optimization methods including Particle Swarm Optimization (PSO), Nelder-Mead Method, and Powell's method. These methods are employed to assign optimal weights to the individual models based on their performance on the aforementioned tasks.
- Recognizing the significance of special characters, such as punctuation keys, contractions, and short and stop words, in distinguishing writing styles, we conduct extensive experiments on both clean and raw (i.e., unclean) datasets. This exploration aims to assess the impact of utilizing these often-removed special characters during pre-processing on the performance of our proposed methods for style change detection.

The rest of the paper is organized as follows. Section 2 provides an overview of the literature. Section 3 provides a detailed description of the tasks tackled in this work. Section 4 describes the proposed methodology. Section 5 provides an overview of the dataset and the experimental setup used in different experiments. Section 6 provides a detailed description of the conducted experiments, experimental results, and key lessons learned. Finally, Section 7 concludes the paper.

2 Related Work

Stylometry aims to analyze authors' style in written documents. It has been studied extensively in the literature [24]. Several interesting works exploring different aspects of the topic [18] have been proposed. Authorship attribution, which relies on SCD techniques, is also one of the key and widely explored aspects of stylometry [27]. In this section, we provide a detailed overview of the literature on SCD with a particular focus on the classification of single and multi-authored documents and author change detection.

Single Vs. Multi-authored Document Classification

The classification of single and multi-authored documents is one of the well-explored tasks of SCD. Several aspects of the tasks, such selection of lexical and deep features, have been already explored in the literature [14]. For instance, Glover et al. [10] analyzed stylometric inconsistencies in multi-authored documents. The authors explored the associated challenges and potential solutions for the detection of stylometric inconsistencies in collaboratively written documents. They also proposed an SCD method that employs different stylometric features, such as word and sentence length, and distribution of short words for detecting authors' boundaries in multi-authored documents. Similarly, Akiva et al. [4] analyzed and separated segments written by the same authors in multi-authored documents using a clustering technique. Several interesting solutions were also proposed for this task in response to PAN-21 shared task [13,28]. For instance, Zhang [30] proposed a BERT-based text classification framework for the classification of documents written by single and multiple authors. Strom [25] also employed the BERT model along with a classifier trained on lexical features including character-based features (e.g., distinct special characters, spaces, punctuation, etc.,), word-based features (e.g., average word length), and sentence-based features (e.g., Part of Speech (PoS) tags, sentence length, etc.,). Moreover, the authors used words function and contracted words in the lexical feature set. Singh et al. [23] also employed lexical features, such as n-grams, POS-Tag, the use of special characters, frequency of words, and the average number of characters per word. A similar set of lexical features is also used by Deibel et al. [7]. The feature set includes sentences as well as word features, such as mean sentence length, mean word length, and function word frequency. One of the solutions proposed for the shared task is also based on Siamese Neural Network architecture with bi-directional LSTM trained on lexical features [17].

Author Change Detection

Author change detection involves the identification of places where the author switches in a collaboratively written document. There are several compelling studies on this task. For instance, Zhang [30] proposed a BERT-based framework that takes pairs of paragraphs from multi-authored documents as input and predicts whether they are written by the same author or not. Nath [17] proposed a Siamese neural network-based solution where each branch of the Siamese network is fed with a separate paragraph. The author conducted experiments under two experimental setups: one with a bidirectional LSTM model with the Siamese network and the other with a bidirectional GRU (Gated Recurrent Unit) network.

Some studies on author change detection also utilize lexical features. For instance, Deibel et al. [7] incorporated a range of lexical features, including mean sentence length, mean word length, and function word frequency at the sentence and work levels, into an LSTM model. Storm [25] also employed lexical features in an ensemble method with BERT embeddings. The lexical features include sentence, word, and word-based features, which are then used to train separate classifiers. Finally, the classification scores of the classifiers are combined in an ensemble method.

In this work, unlike existing solutions, we primarily utilize state-of-the-art transformers, employing multiple pre-trained transformer-based models both individually and jointly in a merit-based late fusion framework. Additionally, we investigate the potential of special characters, which are typically removed during pre-processing in various NLP applications, on the tasks.

3 Tasks Description

This paper mainly focuses on style change detection, which involves the identification of the locations in a multi-authored document where the author switches. This task is further divided into three subsequent subtasks: (i) classification of documents written by single and multiple authors, (ii) identification of the position of changes in documents written by multiple authors, and (iii) identification of all the positions where style changes in documents written by multiple authors. These tasks are described in detail below.

- **Single vs. Multiple Authors Classification** The first step in style change detection is to distinguish between documents written by single authors and those written by multiple authors. Multi-authored documents are written by two or more authors and may contain several switches between the authors. We note that the documents are organized into different paragraphs and in multi-authored documents, individual authors are likely to write separate paragraphs as shown in Figure 1. Figure 1(a) shows an example of a single-authored document, where both paragraphs are written by the same author (i.e., author X). In Figure 1(b), the document consists of two paragraphs written by different authors (i.e., author X and author Y). We note that the multi-authored documents must contain at least two paragraphs. Each sample for the task is labeled either '1' or '0' where '0' represents a single author while '1' represents multi-authored documents.

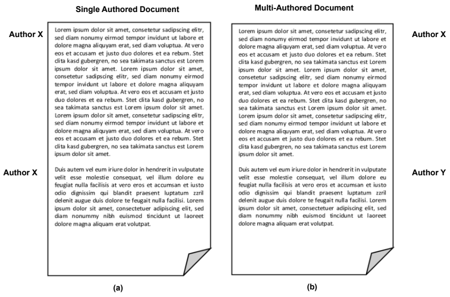

Fig. 1 Single and multi-authored document samples: (a) single-authored document, (b) multi-authored document [28].

- **Style Change Basic (Change Positions)** The second task aims to identify the position where the author switches in a multi-authored document. Similar to task 1, the documents are organized into paragraphs and the proposed solutions predict whether there is a change in author at the boundary of two paragraphs. Figure 2 displays a sample document consisting of three paragraphs written by two authors (i.e., author X and author Y). The red line highlights the position where the author switches (i.e., from X to Y) between the first two paragraphs, while there is no change in the author at the junction of the second and third paragraphs, both of which are written by the author Y. The annotation for the task is provided in terms of '0' and '1' where '0' represents 'no changes' while '1' indicates that the author has been changed at the position.

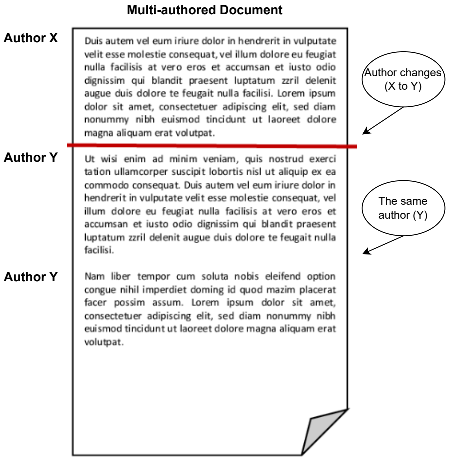

Fig. 2 A sample multi-authored document with a single author switching [28]. The red line indicates the position where the author switches from author X to author Y.

- **Style Change Real-World (Author Attribution)** The third task involves the identification of all positions in a document where the writing style changes (i.e., the author switches between the paragraphs). The proposed solutions should be able to automatically assign a paragraph to each author. In simple words, it's an author recognition task where the proposed system must predict the author of each paragraph. Figure 3 shows an example document from task 3, which is written by three authors, namely author x, y, and author z. The positions where the author switches are represented by a red line. The annotations for this task are provided in terms of $1, 2, 3, ...N$, where N represents the total number of authors contributed to the document.

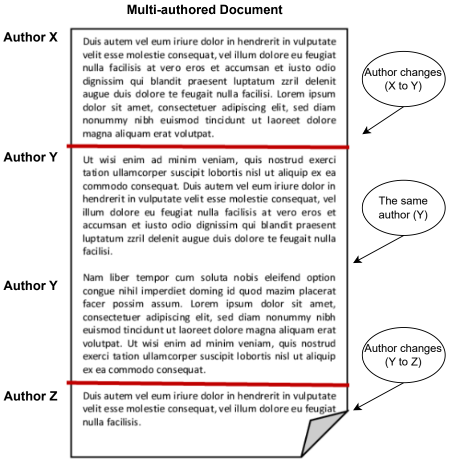
Multi-authored Document
Author X
Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi. Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat.
Author changes (X to Y)
Author Y
Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat. Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis at vero eros et accumsan et iusto odio dignissim qui blandit praesent luptatum zzril delenit augue duis dolore te feugait nulla facilisi.
The same author (Y)
Author Y
Nam liber tempor cum soluta nobis eleifend option congue nihil imperdiet doming id quod mazim placerat facer possim assum. Lorem ipsum dolor sit amet, consectetuer adipiscing elit, sed diam nonummy nibh euismod tincidunt ut laoreet dolore magna aliquam erat volutpat. Ut wisi enim ad minim veniam, quis nostrud exerci tation ullamcorper suscipit lobortis nisl ut aliquip ex ea commodo consequat.
Author changes (Y to Z)
Author Z
Duis autem vel eum iriure dolor in hendrerit in vulputate velit esse molestie consequat, vel illum dolore eu feugiat nulla facilisis.

Fig. 3 A sample multi-authored document with multiple authors switching positions [28]. The red lines indicate the positions where the author switches (i.e., X to Y and Y to Z) while the absence of the red line between paragraph 2 and 3 indicates that the author remains the same.

4 Methodology

We approach all three tasks as a classification problem and use a similar methodology for all of them. Our approach varies mainly in the pre-processing and data preparation phases, where we adopt customized techniques to prepare the data for classification.

As shown in Figure 4, our methodology is composed of three components: (i) preprocessing, (ii) feature extraction and classification through individual models, and (iii) fusion of the scores obtained from individual models. The last two components are common to all the tasks, and we use the same models and algorithms for classification and fusion. The pre-processing component differs slightly for each task, as we employ techniques that address specific challenges associated with the task and the corresponding data. We describe each of these components in detail below.

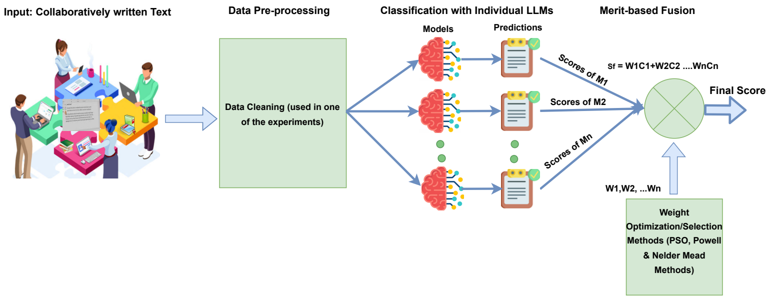
Input: Collaboratively written Text
Data Pre-processing
Classification with Individual LLMs
Merit-based Fusion
Models
Predictions
Data Cleaning (used in one of the experiments)
Scores of M1
Sf = W1C1+W2C2 ....WnCn
Final Score
Scores of M2
Scores of Mn
W1,W2, ...Wn
Weight Optimization/Selection Methods (PSO, Powell & Nelder Mead Methods)

Fig. 4 The proposed methodology.

Pre-processing

In the pre-processing step, we mainly focused on data cleaning, which mainly involved the removal of special characters, such as emojis, stop and short words, and the expansion of contractions. We note that the dataset has already been cleaned by the task organizer by removing URLs, code snapshots, and small paragraphs as detailed in Section 5. Based on our previous experience on the task [27], in this work, we extended the scope of pre-processing by exploring different aspects of the tasks. For example, we tried the removal of different characters one by one and analyzed how the performance varies as detailed in Section 6. Moreover, during our analysis of the data, we observed that in some of the documents, the number of authors and paragraphs was not consistent, which were dropped from the dataset. In addition, we also employed upsampling technique (SMOTE) to balance the dataset. We note that in this work, we did not conduct extensive experiments on the dataset, however, a detailed comparison of results of the proposed methods on tasks with and without data balancing can be found in our earlier work [27].

Feature Extraction and Classification

For text classification in all the tasks, we rely on various transformer-based language models including BERT (Bidirectional Encoder Representations from Transformers) and its multiple variations, such as DistilBERT, ALBERT, Roberta (Robustly Optimized BERT), and XML-Roberta. The selection of these algorithms is made on the basis of their performances in similar text classification applications, such as natural disaster analysis [21], water quality analysis in text [2], and healthcare [1]. A brief overview of these transformers-based models is provided as follows.

- **BERT**: BERT [8] is one of the pre-trained transformer-based models that have been proven very effective in text analysis. It has been trained on a large collection of Wikipedia unlabeled text and can be fine-tuned for text classification in different application domains. One of the key characteristics of BERT includes its ability to extract contextual information/features through bi-directional learning (i.e., left → right and right→ left). Despite being very effective in text analysis BERT also has some limitations. For instance, it is computationally intensive. The literature also reports that the model has not been efficiently trained. To overcome these limitations, several variants of BERT have been introduced in the literature, such as RoBERTa and Distil BERT. Moreover, the original model is available in two different configurations having a different number of layers and attention heads. We employed the version of the model with 12 layers and attention heads and 110 million parameters. We used the Tensorflow implementation of the model with Adam optimizer and binary cross entropy loss function. We trained the model for 5 epochs with a learning rate of $1e-5$.

- **DistilBERT**: DistilBERT [22] is one of the variants of BERT designed to overcome some limitations of the original implementation. It provides a computationally lighter and faster solution for text analysis. The model is based on the concept of knowledge distillation [11], during the pre-trained phase, resulting in a 40% reduction in model size. While the overall architecture remains similar to the original BERT, DistilBERT removes token-type embeddings and the pooling layer, thereby reducing the number of parameters and model size. Despite the significant reduction in the size, DistilBERT retains 97% of the knowledge from the original model. Similar to BERT-base, we used the Tensorflow implementation, employing binary cross-entropy, the Adam optimizer, a learning rate of 1e-5, and training for 5 epochs.
- **RoBERTa**: RoBERTa [16] is another transformer model that emerged as a result of the efforts for improving the BERT model. RoBERTa is built on top of the BERT by making several changes to the training process. For instance, a larger mini-batch size and a higher learning rate have been used in training RoBERTa. Similarly, the model has been trained on a larger dataset for a longer time. More importantly, the BERT's next sentence pre-training objective has also been removed from RoBERTa.
- **XML-RoBERTa**: XLM-RoBERTa is a variant of RoBERTa that is mainly used for multi-lingual text analysis. It is trained on 2.5TB of data containing text from 100 different languages. Though all the text samples in our dataset are in English, XML-RoBERTa could complement the other models in the fusion as demonstrated in similar kinds of applications [2]. We used the same parameter and hyper-parameter settings for RoBERTa (i.e., learning rate = $1e-5$, number epochs = 5, loss function = binary-cross entropy, and Adam optimizer).
- **ALBERT**: ALBERT [15] is another variant of BERT introduced to minimize the size of the original BERT model. To this aim, two different parameter reduction techniques including (i) parameter sharing and (ii) a reduced size of embedding, have been employed resulting in a significant reduction in memory requirements as well as training speed. Moreover, its modeling inter-sentence coherence capabilities also make it a preferred choice for multi-sentence input text. In this work, we used cross-entropy as a loss function with an Adam optimizer using a learning rate of 1e-5 for 5 epochs.

For a fair comparison, we use the same hyperparameters for all the models, which are summarized in Table 1.

Table 1 A summary of the hyper-parameters used in the experiments.

hyper-parameter	Value
Learning Rate	1e-5
Epochs	5
mini-batch size	32
Dropout	0.1
Optimizer/Training Solver	Adam
Cost Function	binary cross-entropy

Fusion

For fusion, we mainly rely on merit-based late fusion where multiple weight selection and optimization methods are employed to combine the classification scores of the individual models using Equation 1.

In the equation, $F(0,1)$ represents the combined classification score while w_n and p_n denote the weight to be assigned to the model and the classification score obtained with the individual (i.e., nth) model, respectively. The weight optimization methods are used to calculate the values of the weights assigned to each model (i.e., $w_1, w_2, w_3, ...w_n$).

$$F(o,1) = w_1 P_1(0,1) + w_2 P_2(o,1) + w_3 P_3(0,1) + ... + w_n P_n(0,1) \tag{1}$$

ALGORITHM 1: Proposed Merit-based Fusion

Input: N: number of models, $w_1, w_2, w_3, ...w_n$: weights obtained through the weight selection/optimization method, $p_1(0,1), p_2(0,1), p_3(0,1), ...p_n(0,1)$: classification posterior probabilities obtained by the individual models for both classes 0 and 1, M: number of samples in the dataset.
Output: $F(0,1)$: combined classification scores, and L_i: final label of the sample i.

```
for $i$ = 1 to $M$ do
    for $j$ = 1 to $N$ do
        $F_i(0,1)$ = $\sum_{j=1}^{N} W_j P_i(0,1)$
        $L_i$ = $max(F_i(0), F_i(1))$
    end for
    Return $L_i$
end for
```

In this work, three different optimization methods namely Particle Swarm Optimization (PSO), NelderMead, and Powell's method are employed to choose proper weights for the classification models. Moreover, for weight selection/optimization, a separate validation set is used for the computation of an accumulative classification error in the fitness function of each method. We note that the same fitness function, as shown in Equation 2, is used for all three methods employed in this work. In the equation, $error$ and A_{acc} represent the classification error and accumulative accuracy computed on the validation set, respectively.

$$error = 1 - A_{acc} \tag{2}$$

Our choice for the weight selection/optimization is mainly based on the proven performance of the methods in similar applications [2]. Some key aspects of these methods are provided below.

- **Particle Swarm Optimization-based Fusion** PSO [12] is one of the widely explored optimization methods and has been employed in a diversified set of applications for different tasks. For instance, it has been used for hyper-parameter optimization of Neural Networks (NNs) [26] and Federated Learning (FL) [20]. The literature also reports its applicability in fusion to select/optimize weights to be assigned to the individual models/classifiers [2]. The algorithm works in several steps starting with a random population of candidate solutions (i.e., different weight combinations). The algorithm then strives to iteratively improve the solution by evaluating the candidate solutions against the given criteria described in its fitness function. During the process, the algorithm keeps track of four parameters including personal and global-best, and the position and velocity of each particle. These parameters are updated at each iteration accordingly until a global minimum is obtained. Despite being very effective in different applications, there are several limitations of PSO. For instance, it is very slow and may fall into the local optimum in high dimensional space. However, in our case, we have a very low-dimensional problem.

- **NelderMead Method-based Fusion** Nelder Mead is also one of the oldest search/optimization algorithms. The name of the method is originated from the names of the researchers (Nelder and Mead) who introduced the algorithm in 1965. Similar to PSO, Nelder Mead has also been widely utilized in a diversified list of applications, and it is available in several ML and statistical analysis libraries [2]. The method is suitable for n-dimensional problems (i.e., $n \geq 0$). Similar to PSO, the method can find the max/min of an objective function, which is based on the accumulative classification error on the validation, resulting in a combination of weights with better classification results. Several implementations of the method are available in different libraries, however, we used SciPy[2] implementation.
- **Powell's Method-based Fusion** Powell's method is also one of the widely explored search and optimization methods. The method is based on the censored maximum-gradients technique. The method is very simple and can be employed in different applications. Similar to PSO, the method solves a problem iteratively by starting with an initial guess (i.e., a combination of weights) and moving towards a minimum of a function. During the process, the algorithm searches and calculates the distance in a good direction. The method also has some limitations, for instance, its assumption of the step size in the direction. In this work, we used SciPy[3] implementation of the method, and our fitness function is defined in equation 2.

5 Dataset and Experimental Setup

Dataset

The proposed solutions are evaluated on a benchmark dataset released for PAN-21 authorship analysis shared task [29]. The dataset covers a total of 16,000 documents taken from the StackExchange network of Q&A sites. For topical consistency, all the documents are taken from sites related to technology, such as code reviews, Computer Science, Data Science, etc. To ensure the quality of the data, several measures were taken by the task organizers. For instance, the text is cleaned by removing URLs, code snippets/screenshots, bullet lists, and documents having modified and altered text. Moreover, the text is organized into paragraphs, and to ensure sufficient text in each sample, paragraphs having less than 100 characters are removed from the documents. Moreover, the dataset contains an equal number of documents per author leading to an equal number of documents written by single, two, three, and four authors. This leads to a data imbalance problem in the first task where 25% documents are single-authored and 75% are multi-authored documents.

[2] https://scipy.org/

[3] https://scipy.org/

We note that the same dataset, which is provided in separate training, validation, and test sets, is used for all the tasks. The training set contains 11,200 (70%) while the test and validation sets are composed of 2,400 (15% each).

Experimental Setup

The main objective of the work is to explore different aspects of stylometry/style change detection. Firstly, we aim to analyze the difficulty level of different tasks involved in stylometry/style change detection by evaluating the performance of the proposed methods in three different tasks. Secondly, we aim to analyze the impact of pre-processing/data cleaning on the performances of the proposed methods in all the tasks tackled in this work. We also want to analyze the performance of different NLP models and fusion methods to jointly employ these models for text classification.

To achieve these goals, we conducted the following experiments using different experimental setups.

- We conducted experiments on three different tasks associated with stylometry/style change detection, each having a different difficulty level and challenges as discussed earlier.
- We conducted experiments on both clean and raw (un-clean) datasets. In the cleaned dataset, we mainly used some basic pre-processing techniques to clean the data. In our previous work [27], we conducted an extensive data cleaning by removing stop words, URLs, short words, separate alphanumeric characters, and expansion of contractions. These techniques are generally used in different NLP applications to improve the performance of text classification algorithms. However, in this work, we conducted an extensive experiment to analyze which of the special characters have more impact on the performance of the models in the application.
- We conducted experiments to evaluate the performances of the individual models and several fusion methods including naive averaging and multiple merit-based late fusion methods.

6 Experimental Results

This section provides the experimental results of the proposed methods on all the tasks. We also provide comparisons against the existing solutions for each task.

Single vs. Multiple Authors Classification

Table 2 provides the experimental results on both clean and raw (i.e., uncleaned) datasets in terms of F1-score. Moreover, the table reports the results of individual models as well as the fusion methods. As can be seen in the table, the performances on the raw data (i.e., the uncleaned dataset) are higher than the cleaned dataset. The maximum F1-scores obtained on the clean and raw datasets are 0.81 and 0.84, respectively, with a gap of around 3%. This gap is lower than what we obtained in our initial work [27]. As discussed earlier, in this work, we expanded the scope of the experiments and explore the impact of different types of characters in the data-cleaning phase. Previously [27], we cleaned the data by removing several special characters, such as usernames, URLs, emojis, stop words, and short words.

Table 2 Experimental results of the proposed methods on both datasets.

Method/Model	Clean Dataset	Raw Dataset
BERT	0.79	0.7946
Albert	0.80	0.7919
Distillbert	0.76	0.7596
Roberta-base	0.8344	.80
XLM-Roberta	0.79	0.83
Simple Fusion	0.770	0.8272
PSO-based Fusion	0.790	0.8486
Nelder-Mead Method-based Fusion	0.810	0.8477
Powell Method-based Fusion	0.811	0.8485

Table 3 Comparison against existing solutions on task 1.

Method	F1-Score
Zhang [30]	0.753
Strom [25]	0.795
Singh et al. [23]	0.634
Deibel et al. [7]	0.621
Nath [17]	0.704
Nelder-mead Method-based Fusion (This work)	0.847
PSO-based Fusion (This work)	0.848
Powell Method-based Fusion (This work)	0.848

Moreover, contractions are also expanded to their original form. However, in this work we observed that the removal of stop and short words has an adverse impact on the performance of the models and thus are kept in the data. The performances of the proposed methods improved significantly by retaining the stop and short words in the data. This improvement in the results indicates that these special characters help in differentiating authors' styles and thus lead to improved results in author style detection.

As far as the evaluation of the individual models and fusion methods is concerned, overall better results are obtained with both versions of RoBERTa (i.e. RoBERTa-base and XML-RoBERTa) while comparable results are obtained with the base BERT and its variants. However, significant improvements in performance are observed when these models are jointly employed in fusion. The improvement in the performance over the best-performing individual model on the cleaned dataset is around 2%, which is much higher compared to the gap on the raw dataset. Moreover, the late fusion methods used in this work can be broadly divided into naive (i.e., simple averaging) and merit-based fusion where weights of the models are obtained through the optimization methods described earlier. Overall, better results are obtained with the merit-based fusion with an improvement of around 2% over the simple averaging.

Table 3 provides the comparison of the proposed fusion methods against the existing solutions on task 1. We note that only our best-performing methods on the raw data are considered for comparison against the existing solutions. As can be seen, the best-performing method obtained an improvement of 5.3% over the best existing solution in the literature. Some potential causes of the improvement include the fusion, choice of NLP models, and retaining of the special characters in the text.

Table 4 Experimental results of the proposed methods on both datasets for task 2.

Method/Model	Clean Dataset	Raw Dataset
BERT	0.70528	0.7423
Albert	0.72224	0.7719
Distillbert	0.688	0.7245
Roberta-base	0.7143	0.7618
XLM-Roberta	0.71294	0.7623
Simple Fusion	0.7368	0.7822
PSO-based Fusion	0.7117	0.7753
Nelder Mead method-based Fusion	0.7117	0.7753
Powell method-based Fusion	0.7224	0.7473

Table 5 Comparison against existing solutions on task 2.

Method	F1-Score
Zhang [30]	0.751
Strom [25]	0.707
Singh et al. [23]	0.657
Deibel et al. [7]	0.669
Nath [17]	0.647
Simple Fusion (This work)	0.782
Nelder-mead Method-based Fusion (This work)	0.775
PSO-based Fusion (This work)	0.775
Powell Method-based Fusion (This work)	0.747

Style Change Basic

Table 4 provides the experimental results on task 2 including the individual models and fusion methods' results on both clean and uncleaned datasets. A similar trend has been observed on task 2, where better results are obtained on raw data compared to the cleaned dataset. However, the difference in the performance of the best-performing methods on both datasets is higher (i.e., around 5%) on task 2, clearly indicating the importance of the special characters in the application.

As far as the performances of the individual models are concerned, there is no clear winner on both cleaned and raw datasets. However, overall XML-RoBERTa and ALBERT have slightly better results compared to the other models on both datasets. We also observed some variations in the performances of the models on clean and raw datasets. For instance, DistilBERT obtained comparable results on the raw dataset, however, it obtained a significantly lower F1 score compared to the other models on the cleaned data. Moreover, similar to task 1, the fusion methods also improved the results compared to the individual models. However, the improvement over the best-performing model is only around 2%. One of the reasons for the less improvement in the result could be the adverse impact of the low-performing model (i.e., DistilBERT).

We also provide comparisons against the existing solutions for task 2. As shown in Table 5, overall, most of our fusion methods outperformed the best-performing existing solution. Our best-performing method on the task (i.e., simple fusion) has an improvement of 3% over the state-of-the-art on task 2. Some potential causes of the improvement include the fusion and retaining of the special characters in the text.

Table 6 Experimental results of the proposed methods on both datasets for task 3.

Method/Model	Clean Dataset	Raw Dataset
BERT	0.5309	0.5244
Albert	0.4998	0.494
Distillbert	0.5261	0.5213
Roberta-base	0.48	0.47
XLM-Roberta	0.5281	0.5159
Simple Fusion	0.5087	0.5164
PSO-based Fusion	0.5535	0.5404
Nelder-mead Method-based Fusion	0.5535	0.5404
Powell Method-based Fusion	0.5453	0.5323

Table 7 Comparison against existing solutions on task 3.

Method	F1-Score
Zhang [30]	0.501
Strom [25]	0.424
Singh et al. [23]	0.432
Deibel et al. [7]	0.263
Nelder-mead Method-based Fusion (This work)	0.5404
PSO-based Fusion (raw-dataset, this work)	0.5404
PSO-based Fusion(clean-dataset, this work)	0.5535
Nelder-mead Method-based Fusion (clean-dataset, this work)	0.5535

Style Change Real-World

Table 6 provides our results on the third task. As can be observed in the table, in contrast to the first two tasks, comparable results are obtained on both clean and raw datasets. The performances of the proposed methods on the cleaned dataset are even better in some cases. For instance, the performances of the PSO and Nelder Mead method-based fusion on the cleaned dataset are better than all the methods on the raw dataset. Moreover, in the individual models, the BERT-base model obtained the highest score while PSO and Nelder Mead-based fusion methods are the best-performing fusion methods on the task.

We also provide a comparison against the existing methods for task 3. As can be observed in Table 7, our best-performing methods PSO and Nelder Mead-based fusion methods have an improvement of around 5% over the best-performing existing solutions.

Lessons Learned

Some key lessons learned from this work can be summarized as follow.

- As indicated by the experimental results, the complexity and the performances of the NLP models vary on the different tasks involved in SCD. The classification of single and multi-authored documents is more straightforward compared to the identification of text segments where the author switches. Similarly, the performance of the proposed methods is significantly lower on the third task involving the identification of multiple locations where the author switches.
- The individual models complement each other when jointly combined in the fusion, which leads to an improvement in the performance.

- The special characters and other words, such as stop and short words, which are generally removed in different NLP applications, play a vital role in the SCD. More, specifically, we observed the removal of stop and short words significantly reduced the performance of the proposed methods.

7 Conclusions

In this paper, we tackled three different tasks of stylometry/style change detection by proposing a merit-based fusion framework to jointly employ multiple NLP algorithms from text classification. In all three tasks, the performance of several transformer-based models and late fusion methods. We also analyzed the impact of pre-processing/cleaning on the performances of the proposed solutions in the tasks. The performance of the proposed methods on all the tasks demonstrates the potential of NLP algorithms in automatic SCD, which can lead to several interesting real-world applications.

In the future, we aim to further explore the change detection by tasks by exploring other relevant subtasks in standard and informal (e.g., social media) texts. We will also explore the potential of these methods in differentiating text written by human authors and chatbots or text generators, such as ChatGPT.

8 Funding and/or Conflicts of interests/Competing interests

The authors declare no conflict of interest.

9 Data Availability

Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study. The experiments are carried out on a publicly available dataset, which is cited in the paper.

References

1. Ahmad, K., Alam, F., Qadir, J., Qolomany, B., Khan, I., Khan, T., Suleman, M., Said, N., Hassan, S.Z., Gul, A., et al.: Global user-level perception of covid-19 contact tracing applications: Data-driven approach using natural language processing. JMIR Formative Research **6**(5), e36,238 (2022)
2. Ahmad, K., Ayub, M., Khan, J., Ahmad, N., Al-Fuqaha, A.: Social media as an instant source of feedback on water quality. IEEE Transactions on Technology and Society (2022)
3. Ahmad, K., Maabreh, M., Ghaly, M., Khan, K., Qadir, J., Al-Fuqaha, A.: Developing future human-centered smart cities: Critical analysis of smart city security, data management, and ethical challenges. Computer Science Review **43**, 100,452 (2022)
4. Akiva, N., Koppel, M.: A generic unsupervised method for decomposing multi-author documents. Journal of the American Society for Information Science and Technology **64**(11), 2256–2264 (2013)
5. Alvi, F., Algafri, H., Alqahtani, N.: Style change detection using discourse markers. CLEF (2022)
6. Bevendorff, J., Chulvi, B., Fersini, E., Heini, A., Kestemont, M., Kredens, K., Mayerl, M., Ortega-Bueno, R., Pkezik, P., Potthast, M., et al.: Overview of pan 2022: Authorship verification, profiling irony and stereotype spreaders, style change detection, and trigger detection. In: European Conference on Information Retrieval, pp. 331–338. Springer (2022)
7. Deibel, R., Löfflad, D.: Style change detection on real-world data using an lstm-powered attribution algorithm. In: CLEF (Working Notes), pp. 1899–1909 (2021)

8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
9. Eke, D.O.: Chatgpt and the rise of generative ai: threat to academic integrity? Journal of Responsible Technology **13**, 100,060 (2023)
10. Glover, A., Hirst, G.: Detecting stylistic inconsistencies in collaborative writing. In: The new writing environment, pp. 147–168. Springer (1996)
11. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015)
12. Kennedy, J., Eberhart, R.: Particle swarm optimization. In: Proceedings of ICNN'95-international conference on neural networks, vol. 4, pp. 1942–1948. IEEE (1995)
13. Kestemont, M., Tschuggnall, M., Stamatatos, E., Daelemans, W., Specht, G., Stein, B., Potthast, M.: 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., pp. 1–25 (2018)
14. Lagutina, K., Lagutina, N., Boychuk, E., Vorontsova, I., Shliakhtina, E., Belyaeva, O., Paramonov, I., Demidov, P.: A survey on stylometric text features. In: 2019 25th Conference of Open Innovations Association (FRUCT), pp. 184–195. IEEE (2019)
15. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: Albert: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942 (2019)
16. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
17. Nath, S.: Style change detection using siamese neural networks. In: CLEF (Working Notes), pp. 2073–2082 (2021)
18. Neal, T., Sundararajan, K., Fatima, A., Yan, Y., Xiang, Y., Woodard, D.: Surveying stylometry techniques and applications. ACM Computing Surveys (CSuR) **50**(6), 1–36 (2017)
19. Pan, Y., Pan, L., Chen, W., Nakov, P., Kan, M.Y., Wang, W.Y.: On the risk of misinformation pollution with large language models. arXiv preprint arXiv:2305.13661 (2023)
20. Qolomany, B., Ahmad, K., Al-Fuqaha, A., Qadir, J.: Particle swarm optimized federated learning for industrial iot and smart city services. In: GLOBECOM 2020-2020 IEEE Global Communications Conference, pp. 1–6. IEEE (2020)
21. Said, N., Ahmad, K., Riegler, M., Pogorelov, K., Hassan, L., Ahmad, N., Conci, N.: Natural disasters detection in social media and satellite imagery: a survey. Multimedia Tools and Applications **78**(22), 31,267–31,302 (2019)
22. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 (2019)
23. Singh, R., Weerasinghe, J., Greenstadt, R.: Writing style change detection on multi-author documents. In: CLEF (Working Notes), pp. 2137–2145 (2021)
24. Stamatatos, E.: A survey of modern authorship attribution methods. Journal of the American Society for information Science and Technology **60**(3), 538–556 (2009)
25. Strøm, E.: Multi-label style change detection by solving a binary classification problem. In: CLEF (Working Notes), pp. 2146–2157 (2021)
26. Wang, Y., Zhang, H., Zhang, G.: cpso-cnn: An efficient pso-based algorithm for fine-tuning hyper-parameters of convolutional neural networks. Swarm and Evolutionary Computation **49**, 114–123 (2019)
27. Zamir, M.T., Ayub, M.A., Khan, J., Ikram, M.J., Ahmad, N., Ahmad, K.: Document provenance and authentication through authorship classification. In: 2023 1st International Conference on Advanced Innovations in Smart Cities (ICAISC), pp. 1–6. IEEE (2023)
28. Zangerle, E., Mayerl, M., Potthast, M., Stein, B.: Overview of the style change detection task at pan (2021)
29. Zangerle, E., Mayerl, M., Tschuggnall, M., Potthast, M., Stein, B.: Pan21 authorship analysis: Style change detection (2021). DOI 10.5281/zenodo.4589145. URL `https://doi.org/10.5281/zenodo.4589145`
30. Zhang, Z.: Style change detection based on writing style similarity