Ensemble-Based Clustering for Writing Style Change Detection in Multi-Authored Textual Documents

Notebook for PAN at CLEF 2022

Shams Alshamasi[1], Mohamed Menai[2]

[1] *Imam Mohammed Ibn Saud Islamic University, College of Computer and Information Science, Computer Science Department, Riyadh, Saudi Arabia.*

[2] *King Saud University, College of Computer and Information Science, Computer Science Department, Riyadh, Saudi Arabia.*

Abstract

Style change detection aims at detecting writing style breaches: the positions at which the writing style changes and authors switch within a multi-authored document. This task has a significant role in forensic linguistics, cybercrime investigation, and intrinsic plagiarism detection. Detecting authors' switch positions require decomposing the text into its authorial components. One of the feasible solutions to achieve this goal is to cluster the textual document into stylistically homogeneous clusters where each cluster includes all text fragments that are similar in writing style and hence are written by the same author. In this paper, we propose a within-document authorship clustering method based on ensemble learning to tackle style change detection in multi-authored documents. The proposed authorship clustering is an unsupervised learning method that does not require training or parameter tuning. The only parameter needed is the number of authors which is estimated by using an ensemble paragraph clustering to accurately capture author distribution at the paragraph level and precisely predict the number of authors. The experimental results obtained on PAN 2022 test dataset show that the proposed method achieved an F1 score of 0.52 to detect style changes between paragraphs and an F1 score of 0.49 to detect authors' switches between sentences. To attribute authors within a document, the method achieved an F1 score of 0.22, a Diarization error rate of 0.57, and a Jaccard error rate of 0.35.

Keywords

Style Change Detection, Writing Style Analysis, Multi-author Analysis, Authorship Attribution, Authorship Clustering

1. Introduction

Style change detection has become an attractive research area related to authorship analysis. It aims to detect text positions at which the writing style changes and authors switch [1]. The rapid increase in cybercrimes and digital text forensics has led to extreme demand for authorship analysis, particularly, authorship attribution that aims at identifying the original author of a given text [2]. Authorship attribution relies on the fact that authors are distinguished by their unique writing style which is characterized by some stylistic features [3] including text readability that measures the text simplicity, clearness, and assesses the reading ease of an author [4], vocabulary richness that measures the diversity of the vocabulary within a given

CLEF 2022 – Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy

sfalshmasi@imamu.edu.sa (S. Alshamasi); menai@ksu.edu.sa (M. Menai)

text [5], and text length that represents the average length of sentences, paragraphs, and words. For example, some authors usually write short, simple, clear, and easy-to-read sentences with simple, short, and common words while others write long sentences with complex and foreign terms.

Traditional authorship attribution focuses on single-authored documents that are labeled with their original author [6]. In this sense, a model could be trained to predict an author for a given anonymous document. Alternatively, modern authorship attribution focuses on multi-authored documents where multiple authors contribute to a single document. Analyzing multi-authored documents for authorship attribution is challenging due to the absence of prior knowledge about the number of collaborative authors, the number of style changes, the authors' style, and the variation of authors' distribution within a document. Thus, analyzing writing style within multi-authored documents requires decomposing the document into its authorial components as a preliminary step for authorship attribution. This could be achieved by identifying the text positions at which writing style changes. In this regard, style change detection is required for improving authorship analysis in a multi-authored document to achieve robust and precise authorship attribution. The advantages of style change detection in improving authorship attribution are extended to cover other potential areas including forensic linguistics for detecting suspicious documents, cybercrime investigation, law enforcement, social media analysis for detecting identity stealing, intrinsic plagiarism detection, and literary research analysis for literary plagiarism detection.

Style change detection has been one of the attractive shared tasks involved in PAN: the series of shared tasks on writing style analysis and digital text forensics, which holds at CLEF conferences [1]. The evaluation results of the annual style change detection competitions from PAN 2017 to PAN 2021 [6, 7, 8, 9, 10] show that style change detection is extremely challenging and has not been adequately resolved yet. The state-of-the-art style change detection systems have some limitations and demonstrate relatively weak performance, particularly due to the large number of extracted features which involve high space complexity and long-running time [11, 12]. Thus, selecting the appropriate stylistic features to discriminate authors' writing style, and choosing the adequate model to precisely identify authorial boundaries are the main challenges to be tackled. Despite these challenges, the awareness of the significant role of style change detection in improving authorship attribution within a multi-authored document motivates for placing more emphasis on developing the appropriate automated style analysis model that is capable of discriminating authors' writing style and identifying authorial boundaries within a document in a more accurate, robust, and cost-effective way.

As evidenced in [13, 14, 15], clustering analysis is beneficial for tackling the style change detection problem in a textual document in which the document is segmented into stylistically similar groups to detect writing style inconsistency between different text fragments. In this paper we propose ensemble-based authorship clustering method for detecting style changes between paragraphs and sentences. The proposed method is based on ensemble clustering that composes of different K-means clustering models to identify the approximate number of authors who contribute to a document. The predicted number of authors is then used to cluster paragraphs or sentences within a document into disjoint clusters where each cluster includes

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

all text fragments written by the same author. Paragraphs or sentences that are belonging to different clusters are predicted to be written by different authors which indicate style changes between them.

The rest of this paper is organized as follows: Section 2 describes the tasks required for PAN 2022 style change detection competition. Section 3 reviews the state-of-the-art style change detection approaches. Section 4 introduces the proposed authorship clustering method. It highlights the proposed stylistic features at both sentence and paragraph levels and describes the proposed clustering method including the proposed ensemble paragraph clustering for identifying the number of authors within a document and the applied clustering methods for solving each required task. The experimental setup is presented in section 5. Section 6 covers the performance evaluation. It outlines the evaluation metrics, highlights the experimental results, and discusses the obtained results. The paper is concluded in section 7.

2. Task Description

Style change detection competitions are established officially by PAN in 2017. Since 2017, PAN has provided researchers and competitors with complete datasets for the annual style change detection competitions. These competitions involved different shared tasks every year.

This year, PAN 2022 [16] style change detection competition aims to detect writing style changes and authors switch between consecutive paragraphs and sentences as well as to attribute paragraphs to their original authors [17]. This competition involves three main tasks:

Task-1: detect authors switch between consecutive paragraphs given a document written by two authors that contains only single style change.

Task-2: given a multi-authored document written by two or more authors the task is to assign a unique author to each paragraph (Authorship attribution at the paragraph level.)

Task-3: given a text document written by two or more authors the task is to identify the text positions at which writing style changes and authors switch between sentences.

Figure 1 illustrates an example of some possible scenarios and the expected output.

As shown in Figure 1, the output of Task-1 and Task-3 is represented as a list of binary values (0 or 1) where 0 indicates no style changes between consecutive paragraphs (Task-1) or consecutive sentences (Task-3), and 1 indicates a style breaches and author switch between consecutive paragraphs (Task-1) or sentences (Task-3). The output of Task-2 is a list of integer numbers that represent the authors attribution result where each number represents a unique author who contributes to a document.

Document A shows an example of authors switch existing between the first and the second paragraph within the document. Document B shows an example of a multi-authored document written by three authors where the first author writes the first paragraph, the second author is assigned to the second and third paragraphs, and the third author is attributed to the fourth paragraph. Document C illustrates an example of a document written by three authors where authors switch between sentences.

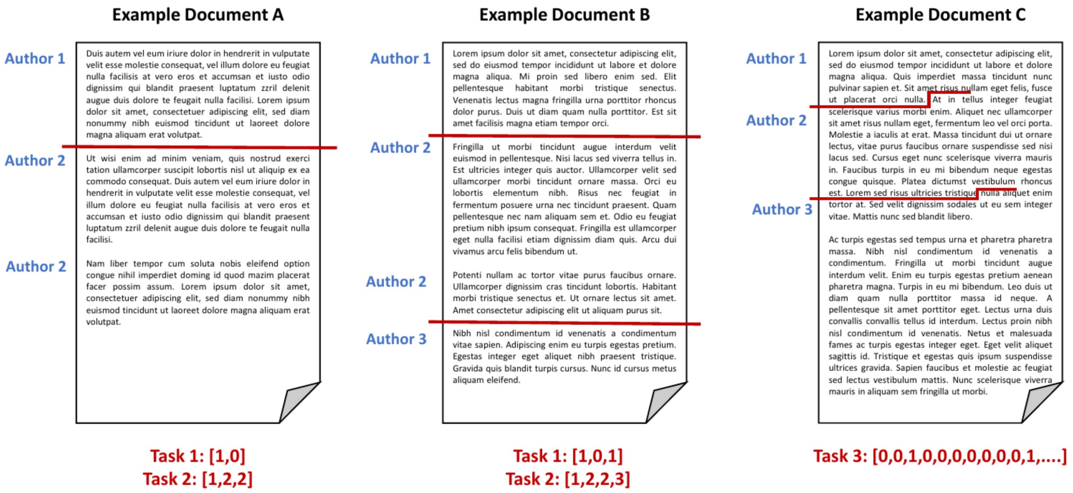

Figure 1: Style Change Detection - Possible Scenarios.
Adapted from https://pan.webis.de/clef22/pan22-web/style-change-detection.html#task

3. Related Works

This section reviews state-of-the-art style change detection approaches that have participated in PAN style change detection competitions from PAN CLEF 2017 to PAN CLEF 2021 [7, 6, 8, 9, 10]. First style change detection approaches were proposed in 2017 including the similarity-based approach which aims to measure the similarity between text fragments and the statistical approach. Khan [18] proposed a threshold-based text segmentation method for segmenting the text into stylistically homogeneous parts by measuring the similarity between adjacent text windows and merges windows that are similar in writing style using a threshold value. A sentence outlier detection method was proposed by Safin and Kuznetsova [19] to measure the similarity between neural sentence embeddings. Wilcoxon Signed-Rank statistical test [20] was proposed by Karas et al. [21] to verify whether two consecutive paragraphs have a significant stylistic differences.

Machine learning approaches have been proposed since 2018 to solve different style change detection tasks. Supervised machine learning methods (binary classification methods) were proposed to discriminate single-authored documents from multi-authored ones. A stacking ensemble classifier based on some lexical and syntactic features was proposed by Zlatkova et al. [22] for classifying documents into single-authored or multi-authored. Safin and Ogaltsov [23] proposed another ensemble classifier to discriminate single-authored documents from multi-authored ones. Strom [24] also proposed a stacking ensemble classifier based on BERT [25] embeddings and features proposed by Zlatkova et al. [22] to verify whether a document is single or multi-authored as well as to detect style changes between paragraphs. The proposed approach uses a recursive method based on the prediction generated by the ensemble to attribute an author to each paragraph. An authorship verification method was proposed by Singh et al. [26] to detect style changes between paragraphs and attribute each paragraph to its original author. A simple and fast divide-and-conquer method was proposed by Khan [27] to measure the similarity between text groups using some lexical and syntactic features to predict whether a document is written by a single or multiple authors.

Unsupervised machine learning approaches (clustering methods) were proposed to identify the number of authors who contribute to writing a document as well as to detect style changes between paragraphs. Window merge clustering and threshold-based clustering using the top 50 frequent terms were proposed by Nath [28] for segmenting the document into stylistically homogenous groups. Castro et al. [29] proposed a non-overlapping B0-maximal clustering algorithm based on some lexical and syntactic features to cluster paragraphs within a document. The method uses a heuristic based on paragraph order to minimize the overlap between the generated paragraph clusters. Zuo et al. [30] proposed a hybrid approach that combines supervised and unsupervised methods to identify the number of authors within a multi-authored document. The hybrid approach first predicts whether a document is written by one or more authors using a feed-forward neural network trained on the TF-IDF document representation. An ensemble clustering of Kmeans and a hierarchical clustering algorithm based on lexical and syntactic features is then used to identify the number of authors at the paragraph level. Another hybrid approach that combines deep features and classification method was proposed by Iyer and Vosoughi [31] to detect style changes between paragraphs using the BERT sentence embeddings in conjunction with a random forest classifier.

Deep learning approaches were proposed to verify whether a document is written by one or more authors. A Character-based Convolutional Neural Network (CNN) was proposed by Schaetti [11] to learn documents' stylistic characteristics for predicting whether a document is written by a single author or multiple authors. A parallel multi-level Recurrent Neural Network (RNN) was proposed by Hosseinia and Mukherjee [12] to verify whether a document is single or multi-authored. The proposed RNN learns the underlying language structure of a document using sentence parse trees. Siamese Neural Network was proposed by Nath [32] to estimate the similarity between all paragraphs within a given document for predicting style breaches at the paragraph level. Multi-layer perceptron and a bidirectional Long-Short Term Memory (LSTM) were proposed by Deibel and Lofflad [33] using word embeddings generated by the FastText model and some lexical and syntactic features to verify whether one or more authors write a document as well as to detect style changes between paragraphs. The prediction made by the LSTM is also used to attribute authors to each paragraph. A similarity-based classification method was proposed by Zhang et al. [34] to predict style change within a textual document. The proposed method uses a fully connected neural network combined with BERT embeddings to predict the similarity between consecutive paragraphs. The predicted similarity is then used to verify whether a document is written by one or more authors, detect style changes between paragraphs as well as assign authors to each paragraph.

Some of the proposed deep learning approaches have some limitations due to the complexity of these models that involve long running time and high space complexity caused by a large number of extracted features as evidenced in [12] which proposed an RNN trained using a large number of extracted parse trees. Moreover, some proposed deep learning models were not trained or tuned deeply due to the lack of training data which caused overfitting that significantly affects the performance and reduces the accuracy, as evidenced in [11].

The review of the state-of-the-art style change detection approaches shows that ensembles of classic machine learning algorithms usually provide good results for detecting style breaches.

The stacking ensemble classifier proposed by Zlatkova et al. [22] was the best performing approach to distinguish single-authored documents from multi-authored ones. The ensemble clustering proposed by Zuo et al.[13] outperforms the existing works in identifying the number of authors within a document. clustering methods have proven their strengths in learning the hidden stylistic patterns for distinguishing multiple authors, particularly with the absence of prior knowledge about the authorship, as evidenced in [14, 13, 29].

4. Proposed Approach

Identifying multiple authors within a multi-authored document requires decomposing it into its authorial components. To achieve this goal, we propose a within-document authorship clustering approach based on ensemble learning to tackle style change detection within a multi-authored document. The proposed approach aims to first identify the number of authors within a document using an ensemble paragraph clustering and then decompose the textual document into its authorial components by clustering the text into stylistically homogeneous groups where each group includes all text fragments (sentences/paragraphs) that are similar in writing style and written by the same author. This method is based on measuring the similarity between consecutive text fragments (sentences or paragraphs) to verify whether they are written by the same author or not. From this point of perspective, the method is also contributed to the task of authorship verification [35]. The proposed method is an unsupervised learning method that does not require training or parameter tuning. The only needed parameter is the number of the author which is predicted by the ensemble paragraph clustering to capture the author distribution at the paragraph level. These are the main strengths of the proposed method to save the time required for training a model, as well as the efforts needed for tuning parameters. This section highlights the selected stylistic features and describes the clustering methods applied to tackle each of the required style change detection tasks.

4.1. Stylistic Features

The task of style change detection this year aims at detecting authors' switches between paragraphs and sentences. Thus, we proposed sets of features at both paragraph and sentence levels. Table 1 outlines the selected features at the sentence level.

We selected lexical and syntactic features due to their relevance to authors' style as they represent authors writing preferences and stylistic choices in comparison with the other possible features such as context-based features and semantic features which are suitable for modeling the topics or representing text context and meaning. Moreover, the extraction of lexical and syntactic features is easy, simple, and usually requires a short running time.

Sentences are usually short and hence the number of features that could be used to characterize authors writing style at the sentence level is limited. We referred to the fact that some authors usually write long and complex sentences while other authors prefer to use short and simple sentences. Moreover, some authors use complex vocabulary (e.g., long words, 2-syllable words, 3-syllable words, etc.), on the other hand, some authors use simple and short words that compose one syllable. As shown in Table 1 , we selected features that describe the length, simplicity, and complexity of sentences including sentence length, average word length per sentence, and

Table 1
Sentence-Level Features

#	Stylistic Features
1	Sentence Length By Characters
2	Sentence Length By Words
3	Average Word Length
4	Average Word Syllable
5	Stopwords Count
6	Function words Count
7	Punctuation Marks Ratio

average word syllable per sentence. Authors differ in terms of the number of stopwords they usually used including the prepositions, pronouns, and determiners. Also, authors differ in the way they use the function words that represent the grammatical and structural relation between content words but don't have an intrinsic meaning on their own [2] such as auxiliary, modals, qualifiers, etc. We proposed stopword count and function words count that represent the total number of stopwords/function words within a sentence. This is due to the short length of the sentence where most stopwords/function words are not existing and hence the existence of individual stopword/function words does not matter but their count does. We utilized the list of stopwords provided by the NLTK library [36] and the list of function words proposed by Zlatkova et al. [22] Punctuation marks could be a distinctive characteristic for identifying authors since some authors usually used a large number of punctuation marks in comparison with others. We proposed the punctuation ratio that represents the total number of punctuation marks to the total number of words per sentence. Table 2 highlights the selected features at the paragraph-level.

Paragraphs are longer than sentences which makes some features more discriminatory when they are extracted at the paragraph level rather than the sentence level. As shown in Table 2, we proposed the readability that measures the simplicity and clearness as well as assesses the reading ease of a given text since authors differ in simplicity and clarity of their writing. We extracted different readability scores including Flesch Reading Ease Score (FRES) [37], Flesch Kincaid Readability Index (FKRI) [38], Automated Readability Index (ARI) [39], Linsear Write Formula (LWF) [40], and difficult words. These readability scores are extracted using textstat python package [3]. Vocabulary richness that measures vocabulary variation and diversity is also proposed to discriminate authors since some authors have rich language and use unique words while others usually have a limited set of common vocabulary. We proposed n-grams including word bigrams and word trigrams since they capture the association and co-occurrence between terms which allows for capturing the frequent phrases preferred by some authors.

[2] https://www.bitgab.com/english-grammar/function-words

[3] Textstat is a python package that provides functions to measure text readability and complexity (https://github.com/shivam5992/textstat)

Table 2
Paragraph-Level Features

#	Stylistic Features
1	Paragraph Length By Characters
2	Paragraph Length By Sentences
3	Average Sentence Length
4	Average Word Length
5	Vocabulary Richness
6	Readability
7	Stopwords TF-IDF
8	Top 50 Frequent Terms
9	Words N-grams (Bigrams and Trigrams)
10	Top 150 Character N-grams (Bigrams and Trigrams)
11	POS Frequency
12	POS N-grams (Bigrams and Trigrams)

Character n-grams are proposed to capture some low-level information. By varying n, character n-grams allow for capturing some punctuation marks, special characters, frequent stopwords, and some short function words. We extracted the top 150 frequent character bigrams and trigrams to reduce the dimensionality of the extracted characters. Finally, since authors differ by the number of different Part-Of-Speech (POS) tag categories (e.g., the number of used nouns, verbs, prepositions, modals, etc.), we propose the pos frequency to measure the frequency of different pos tags categories that appeared in a given text. We utilized a POS dictionary proposed by Zlatkova et al. [22]. POS n-grams represent the syntax of a given text by capturing the location of different POS categories (e.g., whether the adjectival phrase occurs before or after the subject). We proposed POS n-grams due to their ability to capture the grammatical aspects since authors are characterized by their unique way of using the language while writing. For example, some authors write the adverb at the beginning of the sentence while others prefer to write the adverb in the middle of a sentence before the verb.

4.2. Clustering Method

To detect writing style inconsistency between paragraphs and sentences we propose an authorship clustering method. We implemented the authorship clustering method to measure the style similarity between paragraphs and sentences and cluster the paragraphs or sentences into well-separated clusters in which each cluster includes all paragraphs/sentences written by the same author. Each generated cluster represents a unique author, and hence, the number of the generated clusters represents the number of authors who contribute to the document. Based on the problem description, authors can switch between paragraphs (Task-1 and Task-2) but not within a paragraph, and hence each paragraph should be written by exactly one unique author. For Task-3, authors switch between sentences, and hence each sentence should be assigned to exactly one author. Thus, the generated clusters should be well-separated and non-overlapping. For this reason, we selected a partition-based clustering method in which all paragraphs/sentences are clustered into mutually disjoining groups without any overlap between clusters.

Various partition-based clustering algorithms could be used for text clustering. We realized the popularity of the K-means clustering algorithm for authorship clustering. K-means is simple, easy to implement, and usually converges fast. Due to its strengths, we selected the K-means clustering algorithm to tackle the style change detection problem in textual documents. We used a simple K-means algorithm [4] provided by the Scikit-learn library [41] with random centroids initialization and Euclidean distance to measure the similarity between text fragments and cluster centroids. K-means algorithm requires the input parameter k that represents the number of disjoint clusters/partitions to be pre-determined. This was the main challenging step in the proposed approach since the number of clusters that represent the number of collaborative authors is unknown. Hence, the goal first is to automatically estimate the number of authors who contribute to the document. We propose an ensemble paragraph clustering method to find the approximate k value corresponding to the number of authors.

4.3. Ensemble Paragraph Clustering

Detecting authors' switch between sentences is challenging since sentences are usually very short to discriminate authors. Paragraphs are the smallest meaningful text unit that could be used for characterizing authors writing style and capturing authors distribution within a document. From this point of perspective, we propose to identify the number of authors who contribute to writing a document at the paragraph level. The proposed method aims at generating ensemble paragraph clustering models by clustering document paragraphs using different values of possible k. Each generated paragraph clustering model is then evaluated using an optimization score. The best clustering model that optimizes the required score is then selected and the optimal k value that corresponds to the selected model is then used to cluster text within a document. To ensure that the generated clusters are well-separated, the best clustering model to be selected is the model that minimizes the intra-distances between cluster instances and maximizes the inter-distance between different clusters.

Within-Cluster Sum-of-Square Error (WCSSE) that measures the intra-distance is proposed in this paper as an optimization score to assess the generated paragraph clustering models. This is similar to an elbow method that selects the optimal k which optimizes WCSSE by considering the value of WCSSE that shape an elbow. Since the elbow method requires plotting a graph to show the relationship between each possible value of k and the corresponding estimated WCSSE which is hard to be plotted per document, thus we define an optimization score that guarantees to find the optimal k value corresponding to the WCSSE near to the point-shaped an elbow. We define the best WCSSE score to be the minimum value greater than or equal to the average WCSSE estimated from the ensemble paragraph clustering. Figure 2 illustrates the

[4] https://github.com/scikit-learn/scikit-learn/blob/16625450b/sklearn/cluster/$_{k}means.pyL1126$

proposed ensemble paragraph clustering to find the optimal K corresponding to the number of authors.

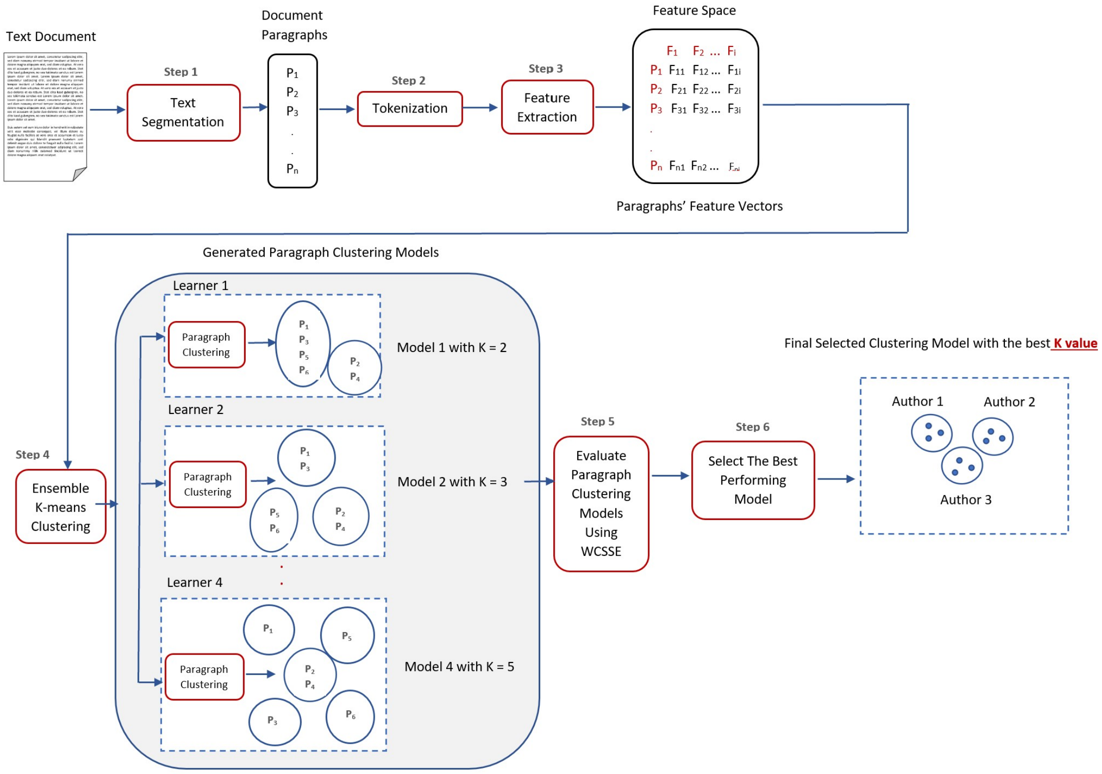

Figure 2: Ensemble Paragraph Clustering

As shown in Figure 2 , the input document is first segmented into paragraphs. Paragraphs features (Table 2) are then extracted from each paragraph resulting in representing each paragraph as a feature vector. Paragraphs feature vectors are then fed into an ensemble paragraph clustering based on a k-means algorithm at which k is ranged from two to five. The minimum value of k is set to two since the minimum number of authors who collaborate in writing a single document is two while the maximum value is set to five based on the description of the input documents which states that documents are written by up to five authors. The generated paragraph clustering models are then evaluated using the WCSSE (inertia). The average WCSSE is then estimated. The optimal K value is the value corresponding to the minimum WCSSE greater than or equal to the average WCSSE. The selected K value at the paragraph level is then used to cluster paragraphs or sentences. Below is a description of how we adapt the proposed ensemble-based authorship clustering method for solving each of the required tasks.

4.4. Clustering Method: Task-1

Task-1 aims to detect style changes between paragraphs in a document written by two authors given that a document contains only a single style change. The number of authors in this task is given, and hence the proposed ensemble paragraph clustering to find the optimal number of authors is not used. To solve this task, the input document is segmented into paragraphs and the paragraphs features (Table 2) are extracted from each paragraph. The paragraph feature vectors are then fed into a K-means clustering algorithm where k is set to two. The similarity between paragraphs' writing style characteristics is measured and the paragraphs are grouped into clusters where each cluster includes paragraphs that are predicted to be written by the same author. The resulted paragraph clusters are then used to detect style changes and authors switch between consecutive paragraphs such that if two consecutive paragraphs are belonging to different clusters this indicates a style change between them.

This method would predict multiple style changes between document paragraphs. To be compatible with the problem specification and detect the only single change, we consider all paragraph pairs that are predicted to have a style change between them. We estimate the cosine similarity between these paragraph pairs and merge all paragraph pairs that are very similar in writing style (have a high cosine similarity). To avoid setting a threshold for merging paragraph pairs, we consider only the style change between paragraph-pair that have the minimum cosine similarity between them since this indicates a high style inconsistency between this paragraph pair while other paragraph pairs are merged in one cluster. This ensures that only a single style change position is detected by the method such that this position refers to the text border between paragraph-pair with the minimum cosine similarity.

4.5. Clustering Method: Task-2

Task-2 aims to attribute each paragraph to its original author given that a document is multi-authored. To attribute authors within a document, the approximate number of actual authors needs to be defined. The proposed ensemble paragraph clustering is used to find the optimal number of authors (Section 4.3). To solve this task, the input document is first segmented into paragraphs and the paragraph features (Table 2) are extracted. The generated paragraph vectors are fed into the ensemble paragraph clustering where k is ranged from two to five. The resulted paragraph clustering models are then evaluated using the WCSSE (inertia) to select the best paragraph clustering model that optimizes the inertia. The selected paragraph clustering model is then used to attribute authors to each paragraph. The resulted paragraph clusters are exploited to perform the authorship attribution within a document such that each resulted paragraph cluster represents a unique author and hence should contain all paragraphs written by that author. The Paragraph cluster labels are exploited to represent the unique authors.

4.6. Clustering Method: Task-3

Task-3 aims at detecting authors' switches between sentences. Before detecting authors switch between sentences, the number of actual authors who contribute to writing the document needs to be estimated which is the main challenge of this task. Sentences are very short to characterize the authors writing style. On the other hand, paragraphs are longer than sentences and hence are more representative for authors distribution within a document. From this point of perspective, we propose to use the ensemble paragraph clustering (Section 4.3) to precisely select the optimal number of authors who contribute to writing the document. The selected K value at the paragraph level is then used with the sentence-level features (Table 1) to cluster document sentences.

The input document is first segmented into paragraphs and the paragraph features are then extracted from each paragraph. The generated paragraph vectors are then fed into the proposed ensemble paragraph clustering to find the optimal k value corresponding to the number of authors. The selected k value at the paragraph level is then used to cluster document sentences. The document is segmented into sentences and the sentence features are extracted. Sentence vectors are then fed into a K-means clustering algorithm with the selected K value to cluster sentences into similar groups. The generated sentence clusters are then used to detect authors' switches between sentences such that consecutive sentences belonging to different clusters have a different writing style and hence are predicted to be written by different authors which indicates style changes and authors' switches between them.

5. Experimental Setup

This section describes the conducted experiments including the used dataset, the applied text pre-processing, feature extraction, and the method applied for parameter selection.

5.1. Dataset

Style change detection 2022 dataset composes of text documents that are constructed using user posts collected from various StackExchange sites and covering different topics [17]. All documents are provided in English and written by up to five distinct authors. Three datasets are provided for each required task including dataset1, dataset2, and dataset3 for solving task1,task2, and task3, respectively. Each dataset is divided into training, validation, and testing sets. The training set contains 70% of the whole dataset while each of the validation and testing sets contains 15% of the whole dataset.

5.2. Experimental Setting

Text documents are not preprocessed by stopwords removal, stemming, or lemmatization since the proposed method is based on stylometric features for discriminating author writing style which characterizes by the use of stopwords, function words, POS tags frequency, etc. The only text preprocessing used to prepare the documents for clustering is the text segmentation into paragraphs or sentences and the text tokenization to represent text as a set of tokens for features extraction. Newline is used as a delimiter as specified in the problem description to segment the document into paragraphs for Task-1 and Task-2, or into sentences for Task-3. Segmented paragraphs or sentences are tokenized using the NLTK tokenizer [36]. Features are extracted from paragraphs and sentences during the running time using the NLTK and Scikit-learn library [41]. The segmented paragraphs and sentences are clustered using the K-means algorithm provided by the Scikit-learn library [5].

The value of the K input parameter is selected during the running time by using the proposed ensemble paragraph clustering method (Section 4.3) for both Task-2 and Task-3. During the

[5] https://github.com/scikit-learn/scikit-learn/blob/16625450b/sklearn/cluster/$_{k}means.pyL1126$

running time, text fragments are clustered and style breaches are detected between consecutive paragraphs (Task-1 and Task-2) or consecutive sentences (Task-3). The detected breaches are then written into a solution file corresponding to each input document. The generated solution files are used to evaluate the performance by comparing the truth or actual breaches provided in the dataset with the predicted breaches written to the solution files. Since the method does not require training, both the provided training and validation datasets are used to evaluate the performance.

6. Performance Evaluation

This section outlines the evaluation metrics and presents the evaluation results.

6.1. Evaluation Metrics

Each task is evaluated independently using a macro-averaged F-score across all documents. Task-2 is evaluated using the macro-averaged F-score combined with two extra metrics including Diarization Error Rate (DER) and Jaccard Error Rate (JER) [17] that measure the text fractions attributed incorrectly [6].

6.2. Results

Foremost, the proposed clustering methods were evaluated using the provided training and validation datasets. The methods were then submitted to the TIRA platform [42] and evaluated using a testing dataset. Table 3 presents the obtained evaluation results from training and validation datasets. Table 4 shows the evaluation results of the proposed model on testing dataset and compares its performance against PAN 2022 style change detection random baseline.

Table 3
Performance Evaluation On Training and Validation Datasets

Dataset	Task-1	Task-2			Task-3
	F1-score	F1-score	DER	JER	F1-score
Training Dataset	0.53	0.21	0.57	0.35	0.50
Validation Dataset	0.54	0.22	0.57	0.36	0.50

[6] https://pan.webis.de/clef22/pan22-web/style-change-detection.html

Table 4
Performance Evaluation On Testing Dataset

Model	Task-1	Task-2			Task-3
	F1-score	F1-score	DER	JER	F1-score
Proposed Model	**0.52**	0.22	0.57	**0.35**	**0.49**
PAN 2022 Random Baseline	0.32	**0.26**	**0.54**	0.40	0.48

6.3. Discussion and Analysis

As shown in Table 3 and Table 4, the performance of the proposed methods is stable and they achieve approximately very close evaluation scores on the three datasets.

The evaluation results on PAN 2022 test dataset (Table 4) show that the most challenging task is the authorship attribution within a multi-authored document (Task-2). The proposed method for tackling authorship attribution at the paragraph level (Task-2) achieves a low f1-score of 0.22 but this obtained f1-score is near to the performance of the PAN 2022 random baseline which has achieved f1-score of 0.26. The low performance of Task-2 is due to the challenges involved in this task since the number of authors is not given and the style of each author is unknown. Moreover, the paragraph length is very short which makes the style discrimination between paragraphs very hard, and hence the performance is affected. However, the evaluation results of the style change detection task this year [17] show that our proposed method for tackling Task-2 achieves the best Diarization Error Rate (DER) and Jaccard Error Rate (JER) in comparison with all the submitted solutions [7]. This shows the strength of the proposed paragraph-level features for attributing paragraphs to their original author as well as the strength of the proposed ensemble-paragraph clustering to estimate the approximate number of authors. The proposed clustering method for Task-1 achieves an F1-score of 0.52 and outperforms the PAN 2022 random baseline for this task. In Task-1, the number of authors is given which is much easier than in Task-2, but still, we are restricted to detecting only one single style change between paragraphs. However, our clustering method for Task-1 would detect multiple style changes between consecutive paragraphs where we tackled this by estimating the cosine similarity between all paragraph pairs that are predicted to have style changes between them to keep only one change between paragraphs that have minimum cosine similarity. The challenges in this task are the short paragraph length and the small number of authors (only two authors) which cause the cosine similarity between paragraphs to be very close and hence the discrimination between paragraph pairs to detect only one change becomes much harder. Although Task-3 is hard to tackle due to the short length of sentences and the small number of features used for characterizing the authors writing style at the sentence level, it achieves an acceptable f1-score of 0.49 which is near to the PAN random baseline performance.

[7] https://www.tira.io/task/pan22-style-change-detection

7. Conclusion

In this paper, we proposed ensemble-based authorship clustering method for tackling style change detection in multi-authored textual documents. The proposed method is implemented in three different ways to tackle each of the required tasks of PAN 2022. We proposed a set of stylistic features to characterize authors' writing style at both paragraph and sentence levels. Moreover, we proposed an ensemble paragraph clustering to approximate the number of authors who contribute to writing a single document. The proposed methods outperform the PAN 2022 random baseline for Task-1 and Task-3. The achieved f1-score for Task-2 is low due to the paragraphs' short length which makes author discrimination at the paragraph level much harder. However, the proposed method for tackling Task-2 achieves the best DER and JER in comparison with all the submitted works.

The findings from this research show that the main challenges in writing style change detection are the selection of the appropriate set of features to discriminate authors writing style which require deep knowledge of stylometry and linguistics, and the choice of an effective method to precisely identify the number of authors with the absence of prior knowledge about the number of authors and their distribution within a document.

In the future, we plan to investigate deep learning combined with the clustering for tackling style change detection by exploiting deep features generated from a model embedding layer and performing the clustering on the generated deep features.

Acknowledgments

This work motivates academic and research interests in authorship analysis and stylometry. All thanks to Professor Mohamed Menai for his endless support and encouragement to accomplish this research.

References

[1] J. Bevendorff, B. Chulvi, E. Fersini, A. Heini, M. Kestemont, K. Kredens, M. Mayerl, R. Ortega-Bueno, P. Pęzik, M. Potthast, 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, Springer, 2022, pp. 331–338.

[2] E. Stamatatos, A survey of modern authorship attribution methods, Journal of the American Society for information Science and Technology 60 (2009) 538–556.

[3] K. Lagutina, N. Lagutina, E. Boychuk, I. Vorontsova, E. Shliakhtina, O. Belyaeva, I. Paramonov, P. Demidov, A survey on stylometric text features, in: 2019 25th Conference of Open Innovations Association (FRUCT), IEEE, 2019, pp. 184–195.

[4] S. Karmakar, Y. Zhu, Visualizing multiple text readability indexes, in: 2010 International Conference on Education and Management Technology, IEEE, 2010, pp. 133–137.

[5] S. Ashraf, H. R. Iqbal, R. M. A. Nawab, Cross-genre author profile prediction using stylometry-based approach., in: CLEF (Working Notes), Citeseer, 2016, pp. 992–999.

[6] 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.

[7] M. Tschuggnall, E. Stamatatos, B. Verhoeven, W. Daelemans, G. Specht, B. Stein, M. Potthast, Overview of the author identification task at pan-2017: style breach detection and author clustering, in: Working Notes Papers of the CLEF 2017 Evaluation Labs/Cappellato, Linda [edit.]; et al., 2017, pp. 1–22.

[8] W. Daelemans, M. Kestemont, E. Manjavacas, M. Potthast, F. Rangel, P. Rosso, G. Specht, E. Stamatatos, B. Stein, M. Tschuggnall, et al., Overview of pan 2019: bots and gender profiling, celebrity profiling, cross-domain authorship attribution and style change detection, in: International conference of the cross-language evaluation forum for european languages, Springer, 2019, pp. 402–416.

[9] J. Bevendorff, B. Ghanem, A. Giachanou, M. Kestemont, E. Manjavacas, I. Markov, M. Mayerl, M. Potthast, F. Rangel, P. Rosso, et al., Overview of pan 2020: Authorship verification, celebrity profiling, profiling fake news spreaders on twitter, and style change detection, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2020, pp. 372–383.

[10] J. Bevendorff, B. Chulvi, G. L. D. L. Peña Sarracén, M. Kestemont, E. Manjavacas, I. Markov, M. Mayerl, M. Potthast, F. Rangel, P. Rosso, et al., Overview of pan 2021: Authorship verification, profiling hate speech spreaders on twitter, and style change detection, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2021, pp. 419–431.

[11] N. Schaetti, Unine at clef 2018: Character-based convolutional neural network for style change detection, Training 2980 (2018) 1490.

[12] M. Hosseinia, A. Mukherjee, A parallel hierarchical attention network for style change detection (2018).

[13] C. Zuo, Y. Zhao, R. Banerjee, Style change detection with feed-forward neural networks., in: CLEF (Working Notes), 2019.

[14] S. Nath, Style change detection by threshold based and window merge clustering methods., in: CLEF (Working Notes), 2019.

[15] D. Castro-Castro, C. A. Rodríguez-Lozada, R. Muñoz, Mixed style feature representation and b-maximal clustering for style change detection., in: CLEF (Working Notes), 2020.

[16] J. Bevendorff, B. Chulvi, E. Fersini, A. Heini, M. Kestemont, K. Kredens, M. Mayerl, R. Ortega-Bueno, P. Pezik, M. Potthast, F. Rangel, P. Rosso, E. Stamatatos, B. Stein, M. Wiegmann, M. Wolska, E. Zangerle, Overview of PAN 2022: Authorship Verification, Profiling Irony and Stereotype Spreaders, and Style Change Detection, in: M. D. E. F. S. C. M. G. P. A. H. M. P. G. F. N. F. Alberto Barron-Cedeno, Giovanni Da San Martino (Ed.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Thirteenth International Conference of the CLEF Association (CLEF 2022), volume 13390 of *Lecture Notes in Computer Science*, Springer, 2022.

[17] E. Zangerle, M. Mayerl, M. Potthast, B. Stein, Overview of the Style Change Detection Task at PAN 2022, in: CLEF 2022 Labs and Workshops, Notebook Papers, CEUR Workshop

Proceedings, 2022.

[18] J. A. Khan, Style breach detection: An unsupervised detection model., in: CLEF (Working Notes), 2017.

[19] K. Safin, R. Kuznetsova, Style breach detection with neural sentence embeddings., in: CLEF (Working Notes), 2017.

[20] R. F. Woolson, Wilcoxon signed-rank test, Wiley encyclopedia of clinical trials (2007) 1–3.

[21] D. Karas, M. Spiewak, P. Sobecki, Opi-jsa at clef 2017: Author clustering and style breach detection., in: CLEF (Working Notes), 2017.

[22] D. Zlatkova, D. Kopev, K. Mitov, A. Atanasov, M. Hardalov, I. Koychev, P. Nakov, An ensemble-rich multi-aspect approach for robust style change detection, CLEF (Working Notes) (2018).

[23] K. Safin, A. Ogaltsov, Detecting a change of style using text statistics, CLEF (Working Notes) (2018).

[24] E. Strøm, Multi-label style change detection by solving a binary classification problem, in: CLEF (Working Notes), 2021.

[25] 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).

[26] R. Singh, J. Weerasinghe, R. Greenstadt, Writing style change detection on multi-author documents, in: CLEF, 2021.

[27] J. A. Khan, A model for style change detection at a glance (2018).

[28] S. Nath, Style change detection by threshold based and window merge clustering methods., in: CLEF (Working Notes), 2019.

[29] D. Castro-Castro, C. A. Rodríguez-Lozada, R. Muñoz, Mixed style feature representation and b-maximal clustering for style change detection., in: CLEF (Working Notes), 2020.

[30] C. Zuo, Y. Zhao, R. Banerjee, Style change detection with feed-forward neural networks., in: CLEF (Working Notes), 2019.

[31] A. Iyer, S. Vosoughi, Style change detection using bert., in: CLEF (Working Notes), 2020.

[32] S. Nath, Style change detection using siamese neural networks, in: CLEF (Working Notes), 2021.

[33] R. Deibel, D. Löfflad, Style change detection on real-world data using lstm-powered attribution algorithm, in: CLEF, 2021.

[34] Z. Zhang, Z. Han, L. Kong, X. Miao, Z. Peng, J. Zeng, H. Cao, J. Zhang, Z. Xiao, X. Peng, Style change detection based on writing style similarity, Training 11 (1970) 17–051.

[35] Efstathios Stamatatos and Mike Kestemont and Krzysztof Kredens and Piotr Pezik and Annina Heini and Janek Bevendorff and Martin Potthast and Benno Stein, Overview of the Authorship Verification Task at PAN 2022, in: CLEF 2022 Labs and Workshops, Notebook Papers, CEUR Workshop Proceedings, 2022.

[36] E. Loper, S. Bird, Nltk: The natural language toolkit, arXiv preprint cs/0205028 (2002).

[37] J. Hartley, Is time up for the flesch measure of reading ease?, Scientometrics 107 (2016) 1523–1526.

[38] S. Karmakar, Y. Zhu, Visualizing multiple text readability indexes, in: 2010 International Conference on Education and Management Technology, IEEE, 2010, pp. 133–137.

[39] R. Senter, E. A. Smith, Automated readability index, Technical Report, Cincinnati Univ OH, 1967.

[40] J. C. Brewer, Measuring text readability using reading level, in: Advanced methodologies and technologies in modern education delivery, IGI Global, 2019, pp. 93–103.

[41] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al., Scikit-learn: Machine learning in python, the Journal of machine Learning research 12 (2011) 2825–2830.

[42] M. Potthast, T. Gollub, M. Wiegmann, B. Stein, TIRA Integrated Research Architecture, in: N. Ferro, C. Peters (Eds.), Information Retrieval Evaluation in a Changing World, The Information Retrieval Series, Springer, Berlin Heidelberg New York, 2019. doi:10.1007/978-3-030-22948-1_5.