International Journal of Scientific Research in Computer Science, Engineering and Information Technology
ISSN : 2456-3307 (www.ijsrcseit.com)
doi : https://doi.org/10.32628/CSEIT228617

An Optimal Feature Set for Stylometry-based Style Change detection at Document and Sentence Level

Vivian Oloo*, Lilian D. Wanzare, Calvins Otieno
Department of Computer Science, Maseno University/Kisumu, Kenya

Article Info

Publication Issue :
Volume 8, Issue 6
November-December-2022

Page Number : 295-313

Article History
Accepted: 12 Nov 2022
Published: 28 Nov 2022

ABSTRACT

Writing style change detection models focus on determining the number of authors of documents with or without known authors. Determining the exact number of authors contributing in writing a document particularly when the authors contribute short texts in form of a sentence is still challenging because of the lack of standardized feature sets able to discriminate between the works of authors. Therefore, the task of identifying the best feature set for all the tasks of the writing style change detection is still considered important. This paper sought to determine the best feature set for the writing style change detection tasks; separating documents with several style changes (multi-authorship) from documents without any style changes (single-authorship), and determining the number and location of style changes in the case of multi-authorship. We performed exploratory research on existing stylometric features to determine the best document level and sentence level features. Document level features were extracted and used to separate single authored from multi-authored documents, while sentence level features were used to answer the question of determining the number of style changes To answer this question, we trained a random forest classifier to rank document level features and sentence level features separately, and applied an ablation test on the top 15 sentence level features using k-means clustering algorithm to confirm the effect of these features on model performance. The study found out that the best document level feature set for separating documents with and without style change was provided by an ensemble of features including number of sentence repetitions (num_sentence_repetitions) as the most determinant feature, 5-grams, 4-grams, Special_character, sentence_begin_lower, sentence_begin_upper, diversity, automated_readability_index, parenthesis_count, first_word_uppercase, lensear_write_formula, dale_chall_readability, difficult_words, type_token_ratio. These were the top ranked features in experiment one. On the other hand, the top fifteen sentence level features based on feature ranks using random forest classifier were diversity, dale_chall_readability grade, check_available_vowel, flesch_kincaid grade, parenthesis_count, colon_count, verbs, bigrams, alphabets, personal pronouns, coordinating conjunctions, interjections, modals, type_token ratio and punctuations_count. Consequently, the optimal feature set for determining the number of style changes in documents was considered based on the results of the ablation study on the top fifteen sentence level features, and was provided by an ensemble of features including personal pronouns, check_available_vowels, punctuations_counts, parenthesis count, coordinating conjunctions and colon count.

Keywords: Writing Style change Detection, Feature Set, Ablation Study, Stylometric Features

I. INTRODUCTION

Writing style change detection, defined as the task of identifying authorship changes in a document by examining the writing styles of individual authors and quantifying the differences in writing styles applied in a document, is steadily gaining momentum thanks to the annual PAN tasks [1], [2]. Traditionally it involved verifying whether a document is written entirely by one known author characterized by a uniform style throughout the document, or if it contains elements of style breaches considered as the existence of new authors [3]–[5]. It has since evolved to include even more difficult tasks of determining the total number of authors involved in documents- of known or unknown authorship, identifying the exact places where authorship switch occurs and mapping the authorship switch with its corresponding authors. The authorship switch may be between documents, sections of document such a paragraph, a sentence group or even a sentence. Recent studies report that the task of writing style change detection is even more challenging as the text length decreases [1], [2], [6].

Writing styles have been defined using stylometric features. An author's writing style is represented by stylometric attributes which are persistent throughout all the works of the author [7]–[9]. These patterns can be quantified to a writing style and be used to identify all the works by the said author, or can be used to establish that sections of a document have similar or different authorial styles. The number of authors in a document can be established by analyzing the writing styles presented in each section of a document for similarities through analysis of an individual's writing style [10], [11]. If two sections of the document yield similar writing styles, then the sections have the same author and vice versa. The use of authorial signature is still at its infancy compared to finger printing, although state-of-the art studies indicate their effectiveness in identifying authors [2], [12].

Numerous stylometric features exist in literature which have been used to model the writing styles of authors. These features are categorisable into three, four or five categories. [13] categorizes stylometric features into four categories namely lexical, syntactic, structural and content-specific features. [9], [14], [15] divides the features into five categories: lexical, character, syntactic, structural and context-based features. Further still other studies classify these features in just three groups, namely word-based features, syntactic and content-specific features [16]. Following the previous work [7], [9], [14], [15], [17] we define five categories of stylometric features namely: lexical, character, syntactic, structural and context-based features and group features into these categories.

These features have been applied in writing style change detection studies with varying effects. For instance, most studies report the effectiveness of lexical features in writing style change detection in datasets with homogenous topics [11], [12]. On the other hand, content features are more useful if the dataset contains different topics.

Structural features such as readability which measures the ease of reading a text have been used in previous studies as features. They assess the clarity and simplicity together with the ease of reading a given text. These features can be used to distinguish between different styles since authors differ in simplicity and clarity of their writing. The different readability measures are defined and used in literature. For example, Flesch Reading Ease Score which measures the ease of reading a piece of text. It gives texts scores ranging from 0-100%, with a score of between 70% to 80% , indicating grade eight level. Grade eight level means that an average adult can read the text fairly easily [18]. Flesch Kincaid Readability grade level is a formula which assesses the approximate reading grade level of a text. It converts the flesch kincaid reading ease score to the reading grade level such that a flesch Kincaid level of 8 means that the reader requires grade 8 and above to understand the text [18], [19].

Automated Readability Index measures the United States grade level required to read a text. It counts the number of characters in a text inorder to determine the grade level, such that the higher the number of characters the more difficult a word is to read [4]. Lensear Write Formula is a textbased formula scoring monosyllabic words and strong verbs. A score between 70% to 80%, is favourable for an adult reader while scores below 70% may be considered hard for an average reader [18], and difficult words [4], have yielded promising results even in shortlength documents. SMOG Index on the other hard determines the grade level required to understand the text. Moreover, in very small datasets, character features tend to be more effective [20].

State of the art studies report that pre-trained BERT models could be more effective than stylometric features in most writing style change detection tasks in small datasets[1]. However, this study believes that stylometric features can still compete favorably with pre-training if an optimal feature set for each task is found and standardized. Previous researchers [6], [19], [21] have used various combinations of features sets for writing style detection with promising results but none has so far looked at the standardizing the feature sets to find the most optimal feature set for this task. This study seeks to find the optimal features sets for the task to standardize the features for the task and improve model performance to be at per with BERT like models.

The rest of this chapter is organized as follows: Section A. outlines the background information. Section B. describes the problem statement and section C. presents an analysis of related work.

A. Background Information

Determining the best feature set for a machine learning algorithm remains an important task in authorship studies. The presence of a rich feature set applicable to these tasks is advantageous to researchers as they have the flexibility of selecting which features to use for their models [7]. However, the catch is that the purity of these models rely heavily on the features used. As such to achieve high performance features must be selected that have the right attributes to model authors writing styles, and be able to distinguish between works of different authors. In addition, the diversity of the features used might also thwart attempts to compare performances of different models thereby necessitating standardization of optimal feature sets for the various tasks of authorship studies [22], [23].

This study adopts the feature categorization by [7], [8], [14] which groups stylometric features into five categories namely lexical, syntactic, character, and structural and context features. Lexical features are meant to indicate the preference of a user for a certain group of words or symbols [9]. They can be extracted by dividing the text into tokens where a token can be a word or character. Examples of lexical features include bag of words, word n-grams, vocabulary richness, and most frequent words among others. These features are the most commonly used in writing style change detection studies because of their ability to be used across different languages [14]. Character level features include the characters, comprehending uppercase, lower case, vowels, white spaces, digits, special characters, alphabets, symbols representing the mood of the author and character n-grams. These features are tolerant to typing errors including grammatical errors and misuse of punctuations [20].

Syntactic features can be defined as context-free features and are therefore suitable for studies cutting across different topics [10], [24], [25]. However, they are language-dependent and their use relies on the availability of a syntactic parser. These features include Part of Speech (POS) words and punctuations. The POS feature consists of tagging a word on the basis of its context and it can be classified as verbs, prepositions, contractions, modals, interjections, adverbs, adjectives, nouns, pronouns, conjunctions among others [12], [26]. Structural features are used to capture the overall characteristics of the organization and the format of a text. They can be defined at three levels; document, paragraph and at the technical structure of the document [15], [24]. They include number of sentences in a paragraph or document, average number of words, characters, mean sentence length, average number of sentences beginning with upper and lower cases among others [15]. Context features on the other hand check for keywords signifying the existence of different topics or context. These features are only significant if the dataset contains documents of varying topics and authors are determined based on the topics. Otherwise in homogenous datasets the use of context features does not add value to performance [20].

This paper focuses on determining the right features for the writing style change detection at the document level and the sentence level. The goal of the writing style change detection is to determine the exact number of authors collaborating in writing a document [2]. Different tasks have been investigated under writing style change detection with notable increase in complexity. The fundamental task being that of checking whether a document is single or multi-authored to the complex task of determining locations of authorship switches and the number of authors collaborating in writing a document. [14] conducted an extensive review of existing work on the task of writing style change detection. The review found out writing style change detection is still challenging although there is an upward trend in performance of the state-of-the-art studies. All these tasks rely on document analysis for the existence of different writing styles in the document. The existence of two or more styles signifies multi-authorship while a single style is single authored [5], [21], [27]. This is a new area of research if the number of studies in this area are anything to go by.

Few studies exist for the writing style change detection using stylometric features and machine learning algorithms. For such studies selecting the right features to use remains an important exercise [28]. Literature boasts of a number of feature selection methods ranging from manual to automatic, however most writing style change detection studies built from the conventional authorship verification studies by enhancing the models or using expanded or new features sets [18], [19], [21]. Particular features which yielded promising results in previous studies were considered in recent studies without determining whether they were the most appropriate for those studies [1], [19]. Whereas this approach of feature engineering has proved to be efficient to some extent, they assume that similar stylometric features may have the same effect in all authorship studies. In addition, the same features may have different effects depending on whether they are extracted at the document or at the sentence level. Moreover, contributions of the other features not considered in those studies but which may end up having positive contributions to the performance of the model for a different task is also ignored. Experimental methods have also been have used to engineer features for the writing style change detection [18].

A summary of existing stylometric features and their categories is presented in table 1 below

Table 1: Most commonly used features in writing style change detection adopted from Oloo et al., (2022)

CATEGORY	FEATURES	REFERENCES
Lexical	**Word Level features**	
	Word n-gram	[25], [29], [30]
	Word frequencies	[25], [26]
	Vocabulary richness	[31]
	Stop words count	[4], [26]
	Number of difficult words	[18]
	Word length, total number of words	[6], [32]
	Average word length	[4], [31]
	most frequent words	[3]
	Average word syllable	[4]
	Word pair frequencies	[33]
	Type_token ratio	[34]
	Duplicate words	[18], [19]
	Most frequent terms	[30]
	Sentence Level features	
	Duplicate sentences	[21]
	Sentence length	[4]
	Number of sentences starting with lower case letters	[3]
	Total number of all-uppercase words in a sentence, Number of sentences starting with capital letters	[6], [32]
	Total number of misspelt words	[19]
	Total number of words in a sentence	[4]
Character Level	Special characters such as , Digits, Alphabets, White spaces, Emojis	[35]
	Character n-grams	[3], [30]
	n-gram count	[31], [35]
	Tabs count	[35]
	Special character frequencies	[31]

	Total number of uppercase letter	[35]
	Character frequencies	[33]
	Total number of special characters	[31]
	Most frequent character n-grams	[30]
	First word uppercase	[35]
Syntactic Features	Punctuations such as single quotes, commas, periods, colons, semi-colons, question marks, exclamation marks, and special marks based on Unicode format.	[33]
	Part of Speech words (POS) including common words such as nouns, pronouns, prepositions, adjectives, interjections, conjuctions, verbs, adverbs contractions, determiners, modals etc.	[4], [18], [19]
Context Features	Key words, Interest groups, special activities	[17]
Structural Features	linsear_write_formula, Flesch_kincaid_grade, Diversity, Dale_chale_readability, Automated_readability index	[18], [19]
	special character ratios, ratios of tabs, mean sentence length, average number of words, ratio of uppercase letters	[35]
	average number of characters, average number of sentences beginning with uppercase, average number of sentences beginning with lower case	[6], [32]
	ratio of interrogative sentences	[35]

B. Problem Statement

Writing style change detection models focus on determining the number of authors of a document with or without known authors. The performance of these models rely majorly on the machine learning algorithm and the features sets used. However, the lack of an optimized feature set applicable to writing style change detection tasks at the sentence level may thwart the application of these models in determining the exact number of authors contributing in writing a document particularly when the authors contribute short texts in form of a sentence. Better still the performance of these models may improve in terms of their purity and running time if only the most determinant features in the given dataset are used. Therefore, there is need for engineering features able to separate single documents containing no style changes from those with style changes, and features which can be used to determine the number of style changes in documents with improved results.

C. Related Works

Feature engineering remains an important task for the machine learning-based writing style change detection studies. Determining a suitable feature set that is able to model an author's writing style has a bearing on the performance of these models. It is an important task because it reduces overfitting on the dataset, improves the purity of the style change detection model and reduces computational costs.

Existing feature engineering methods can be grouped into three categories; manual feature selection, reviewing of features and experimental methods [19], [21], [36].

Manual feature selection method was the most commonly used method in early authorship studies to engineer features. This method involves manually analyzing the dataset for attributes capable of discriminating between different styles. For instance, [37] proposed a method based on most frequent features to determine the number of authors in multi-authored documents. [9]and [7] manually selected features for the authorship verification studies on short text. Manually engineered features tend to produce improved performance although it is a tedious process and may be limited if the document length is rather short such as in a sentence. Few features engineered manually by analyzing datasets for determinant features have yielded very promising results. For instance, the study by [21] sought to determine the number of authors in a document using a set of features known to produce good results in the previous studies. They proposed to use an ensemble of three clustering algorithms to determine the number of authors in multi-authored documents. Their method yielded promising performance, however, they observed improvements in model performance when duplicate sentences, a feature they identified after document analysis was used.

Most recent studies in writing style change detection have selected the features for their proposed methods by reviewing features from previous work i.e. using features and feature sets which produced promising results in previous related studies [12], [19], [31]. In this method features or subset of features which produced better results compared to the others for the same or related tasks are picked wholesomely or expanded. For instance, [19] used a feature set based on the winning submission of the previous years' PAN competition to determine the number of authors in multi-authored documents. While [38], developed a method based on google's BERT embeddings which had been shown to produce the best results in style change detection studies. Moreover, other studies would pick the best performing features in previous and add other new features to obtain a feature set for their study. Whereas this method can be used to validate the importance of certain features, it may ignore certain not commonly used features but which may turn out to be important for the task at hand.

Few studies have used experimental feature selection methods such as chi-square, Information gain, computing frequencies and entropy among others to select the best feature set for their studies. [39] performed feature selection experiments involving thirty-nine different types of textual measurements mostly used in authorship attribution studies. His experiments, which were performed using the Chi-squared test on the Telegraph Columnist corpus, concluded that the combination of word and punctuation mark profiles are effective features for representing authors. Similarly, [40] carried out an exploration of 166 features used for authorship attribution including commonly used stylistic features and several others intended to capture emotional tone. He reported that fifteen features, including punctuation marks, pronouns, fog index and average sentence length to be the most useful.

Other experimental feature engineering methods include computing feature frequencies to identify the right features to use. [25] sought to determine the best features for a classification task. They extracted the most frequent feature types with the assumption that these features will be the most determinant for this task. To select the best feature set to use, their method applied ranking of feature types by frequency and the top feature types were considered the most determinant. Similarly, [41] proposed an approach for automatic authorship verification for cross-genre and cross-topic datasets written in four languages- Dutch, English, Greek and Spain. They employed a random forest classifier to choose the most important specific features by computing feature importance scores. A total of seventeen specific features were selected from the main features including punctuation, sentence length, vocabulary, N-gram, Parts-of-Speech (POS).

The study selected the following set of word and style-based features for their model: total number of punctuations, ratio of specific punctuations, long sentences/ short sentences ratio, vocabulary strength, N-gram difference, POS frequencies, POS sequence frequencies and starting POS frequency.

[36] carried out a feature extraction and selection for the Tamil language using decision trees for authorship identification. To validate their proposed method, they experimented with Support Vector Machines, C4.5 algorithms, and Classification Based Associations (CBA). This study reported higher performance when decision trees were used compared to other methods. Computing frequencies is the most widely used method in previous studies to engineer features partly because the most frequent features in a dataset may also be the most determinant in distinguishing between writing styles in documents.

Different machine learning methods can be used for the experimental feature selection. Machine learning algorithms such as tree-based algorithms, support vector machines, classification based associations, C4.5 algorithms among others have been used to rank features based on computations of feature importance scores. In this method the most important features are the highest scoring features and vice versa. [25] Used feature ranking by frequency to select the best features to use in classification tasks. They considered the top features to be the most determinant features for the task. [36] used decision trees to select the best features for an authorship identification task on Tamil language. They experimented their proposed approach using Support Vector Machines, C4.5 algorithms, and Classification Based Associations (CBA). This study reported higher performance when decision trees were used compared to other methods.

Tree-based methods such as random forest and decision trees yield acceptable results when used to engineer features. Such classifiers are fast to train and easy to evaluate and interrupt. Moreover, they are non-parametric and for the very reason they are not affected by outliers. The main shortcoming is that they easily overfit but that's where ensemble methods like Random Forest come in. Random forest is known to have lower classification errors and better F-scores than decision trees. In addition, they train faster and are easy to understand compared to decision trees. [41] proposed an approach for automatic authorship verification for cross-genre and cross-topic datasets written in four languages- Dutch, English, Greek and Spain. They employed a random forest classifier to choose the most important specific features by computing feature importance scores. A total of seventeen specific features were selected from the main features including punctuation, sentence length, vocabulary, N-gram, Parts-of-Speech (POS). The study selected the following set of word and style-based features for their model: total number of punctuations, ratio of specific punctuations, long sentences/ short sentences ratio, vocabulary strength, N-gram difference, POS frequencies, POS sequence frequencies and starting POS frequency.

Some studies have also applied the use of ablation studies commonly used in the field of medicine to determine contributions of each feature category in different datasets. [20] sought to determine the contributions of three feature groups; style features, content and hybrid features on different datasets. They experimented with four widely used datasets in authorship attribution; CAT 10, CAT 50, JUDGEMENT and IMDb62 datasets. Using logistic regression and a Feed Forward Neural Network on the first 100 common n-grams, the study found out that style-based features were more effective for datasets where authors discuss similar topics. On the other hand, content features showed usefulness in datasets with dis-similarity in topics. Ablation tests are the most reliable tests for determining feature importance since they compute the actual feature contribution to the model. Ablation study yields the most promising results compared to feature ranking used [42].

This study also uses feature rankings based on importance score computations and random forest classifier for selecting the best features for separating documents with style change from those without, and to determine the number of style changes in documents [25], [36], [41]. It differs from existing studies in that it first ranks all the features based on their importance scores followed by ablation study to determine feature contributions for determining the number of style changes in documents.

II. METHODS AND MATERIAL

To determine an optimal feature set for the writing style change detection task we employed exploratory research design to engineer features. First we looked at all the features from literature which have been employed in writing style change detection and other related studies such as authorship verification involving shorttext length (see table 1, section A). We adopt the definition of shorttext length as a document containing not more than five hundred characters as per [7]. Only authorship verification studies involving short text length were considered in this study since they mirror the task of writing style change detection. We discuss the methods and materials (methodology) by looking at document pre-processing (section A), dataset used (section B), feature extraction (section c) that looks extraction techniques for the document level and sentence level features, and feature selection (section C) that explain the experiments conducted for feature section.

The rest of the section is organized as follows. Section A presents the document pre-processing carried out on the data. Section B highlights the dataset used. Section C describes the feature extraction and section D. describes the feature selection process.

A. Document Pre-processing

We did not conduct rigorous document processing on the data as this is deemed to remove certain features which can be important for the study (Brocardo et al., 2015). However, we still cleaned the data by removing some frequent phrases which carry little or no linguistic style contributions. These include typical URLs and technical specifications such as “OSX 10.11.2.” Contractions which are a unique type of word that combines two or more other words in a shortened form, usually with an apostrophe were used as a feature in this study. They take words that usually go together, like can not or I have, and then remove certain letters to shorten them and make other words, like can't or I've. Since contractions form part of NLTK’s stopwords library for the English language, we first expanded all of the contractions before removing stop words so that they could be extracted later. Other fundamental preprocessing performed on the data include lemmatization and tokenization.

The Dataset

Authors	1	2	3	4	5
Training	1273	325	313	328	307
Validation	636	179	152	160	145

Table 2: Distributions of authors in the training dataset.

This study used the training dataset for the Pan at CLEF 2019. This dataset contains 2546 documents and a separate dataset for validation consisting of 1272 documents. All the documents were written in English and covered various topics. For each document gold-standard labels showing the number of authors and the annotations marking the authors of each section of the document was provided.

This study used the train: validate: test ratios as provided for in the dataset.

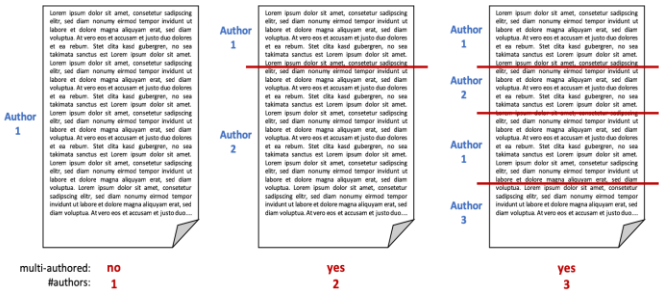


Figure 1: Gold-labels illustrating the different scenarios in the study and the expected outputs. Adapted from https://pan.webis.de/clef19/pan19-web/style-change-detection.html#task.

From figure 1 above, the first scenario is where there is only one style throughout the document i.e no style changes depicting a single author. The second and third images are cases of multi-authors depicted by one style change in the second paragraph and three style changes respectively. Note the red line indicates the exact position where the style change occurs.

After analyzing the data, it was realized from the gold-standard labels that half of the documents in the training dataset were single authored while the remaining half were multi-authored, with the number of authors range from two to five authors as shown in table 2 below.

C. Feature Extraction

Various features were extracted categorized under lexical, syntactic, structural, character and content features. Features were extracted both at the document level and the sentence level. For the task of identifying whether a document has style change or not we extracted features at the document level. While determining the actual number of style changes in a document, we extracted only sentence level features. The following features were extracted:

Lexical: We extracted lexical features at the word, sentence and character level. We use as features proportions of various types of lexical elements in the document. Specifically, we count the total number of occurrences of each lexical element and divide by the total number of elements in a document. NLTK's tf-idf vectorizer was used to extract lexical features. The features included eleven word-based features, seven sentence level features and thirteen character-based features. In general, a total of 32 lexical features at word, sentence and character features were used (see table 1 section A).

Syntactic features: These include POS words and punctuations. POS words were extracted using NLTK's POS tagger was to extract function words such as Part of Speech words (POS) including common words such as nouns, pronouns, prepositions, adjectives, interjections, conjunctions, verbs, adverbs, contractions, determiners and modals. Punctuations were extracted using NLTK's tf-idf vectorizer. The following punctuations and special symbols were extracted: Punctuations such as single quotes, commas, periods, colons, semi-colons, question marks, exclamation marks, and special marks based on Unicode format. A total of nineteen (19) syntactic features were extracted and used for the study.

Structural Features: these features mainly consisted of readability scores and other features such as average word length and mean sentence length among others. This study used Textstat Pythons package to compute and extract the following readability features: lensear_write_formula [19], flesch_kincaid_grade [18], diversity [19], automated readability index [18], dale_chall_readability [18,19] SMOG grade [18,19], Coleman-Liau index [18,19], difficult words [18,19], Gunning fog [18,19]. The rest of the structural features by were manually extracted simply counting the total number of occurrences and dividing by the total number of words in the document. These features include ratios of tabs, special character ratios, ratio of uppercase letters, average number of words, mean sentence length, average number of characters, average number of sentences beginning with uppercase, average number of sentences beginning with lowercase and ratio of interrogative sentences. In general, the study extracted seventeen (17) structural features.

D. Feature Selection

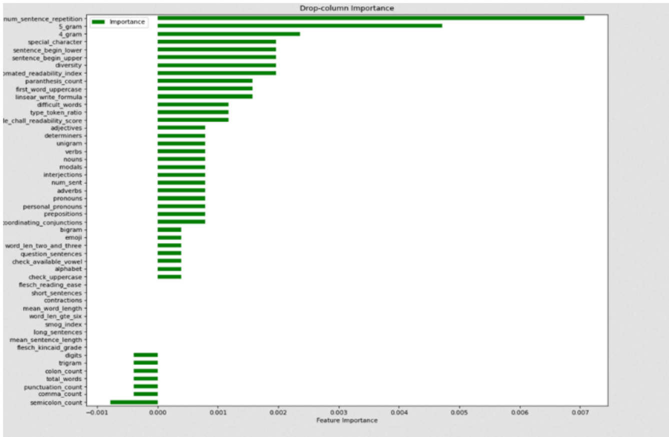
Three experiments were performed. The first experiment was to identify determinant features able to distinguish between documents with and without style change using document level features. The second experiment was designed to identify the most determinant sentence level features for identifying the number of style changes and third experiment was to determine the best feature set for determining the number of style changes in a document by computing the effect of each feature on the model performance.

I) Experiment I

In experiment I, we train a random forest classifier on all the features identified in literature as listed in table 1. For this experiment we use document level features where we calculate the feature importance score for each feature. We use the drop-column importance which is considered to yield the most accurate feature importance. In this method one completely deletes a column from the dataset, retrains the model, and checks how much it affects performance. The idea is to get a baseline performance score as with permutation importance but then drop a column entirely, retrain the model, and re-compute the performance score. The importance value of a feature is the difference between the baseline and the score from the model missing that feature. This strategy answers the question of how important a feature is to overall model performance even more directly than the permutation importance strategy. Although this method is computationally expensive, it is the best method of extracting important features as shown by previous studies [41], [43]. The output of this experiment is a table indicating features and their importance scores ranked in order of increasing importance. Feature importance is basically how much the feature is used in each tree of the forest, and it is computed as the (normalized) total reduction of the criterion brought by that feature. The top fifteen features based on their rankings are picked as the most determinant.

II) Experiment 2:

Experiment II was used to identify the most determinant sentence level features for determining the number of style changes in documents, we train a random forest classifier on all features extracted at the sentence level and compute feature importance scores. The input to the algorithm is a list of labels of all the sentence level features. We use the drop-column importance which is considered to yield the most accurate feature importance like in experiment one. In this method one completely deletes a column from the dataset, retrains the model, and checks how much it affects performance. The output of this experiment was sentence level feature rankings based on their importance scores.



Experiment 3:

In experiment 3 we conducted an ablation test on the top fifteen sentence level features from the second experiment to determine their effect on the overall clustering model performance. The aim was to answer the question 'what is the optimal feature set for determining the number of style changes at the sentence level? To begin this experiment, we created feature vectors of the selected features from the second experiment for all the documents in the dataset. The features were represented as a numpy array, this will be an m x n matrix, depending on the number of sentences denoted as m and n for the number of features. Then we flatten the numpy array to 1 dimension. Because some documents were shorter than the others, we do padding by adding zeros at the end of the list to the shorter documents to ensure all documents have the same dimension.

We then train k-means clustering, setting the value of 1≤k≤5. We start with the entire feature set and iterate through each document in the dataset. We test with different values of k-clusters up to k≤5 because the highest number of authors was set at 5 in the ground truth labels. We then calculated silhouette_score of each cluster and picked the cluster with the best score in the algorithm. To calculate model performance, we use Ordinal Classification Index measure which measures the error of prediction[44]. The effect of each feature was then computed as the difference between the OCI value when the feature was included and when it was not. The output is a table showing contributions of each feature to the performance of the model measured in OCI values. The higher the influence value the more important the feature is to model performance.

In this section we discuss the results of the feature selection experiments. Results of the first experiment which sought to determine the best document level features using Random Forests together with the results of determining the most determinant features at the sentence level are discussed. In addition, we discuss the results of the ablation test.

A. Experiment I

This experiment sought to rank document level features based on the computation of features importance scores using Random Forest. The

aim of this experiment was to determine the best document features for separating documents with style changes from documents without style change as indicated.

The results of the first experiment are shown in FIg 1. Fig 1 presents the feature importance scores of all identified features calculated using Random Forests drop-column importance method. The y-axis represents the most commonly used document level features as identified from literature while the x-axis gives feature importance. Features with zero importance scores have less significance on model performance while those with higher scores have higher significance on model performance.

Results show that number of sentence repetitions (num_sentence_repetitions) ranked highest as the most determinant feature with an importance score of 0.007. 5-grams followed with an importance score of 0.005 followed by 4-grams at an importance of 0.003. special character, sentence_begin_lower, sentence_begin_upper, diversity and automated_readability_index also produced positive importance scores at 0.0025. other features which yielded above zero importance scores were parenthesis_count, first_word_uppercase, lensear_write_formula at a score of 0.002, difficult_words, type_token_ratio, dale_chall_readability_score at an importance score of 0.0015 and adjectives, determiners, unigram, verbs, nouns, modals, interjections, number_sentences, adverbs, pronouns, personal_pronouns and coordinating conjuctions at a score of 0.001. features such as bigrams, emojis, word_len_two_and_three, question_sentences, and check_available_vowels yielded a score of 0.0015.

Some features yielded feature importance scores of zero. These features were considered as not having any significance to the model performance. In other words, using these features contributes no information to the model performance. In the study the features which yielded zero importance scores were flesch_reading_ease, short_sentences, ontractions, mean_word_length, word_len_gte_six, SMOG index, long_sentences, mean_sentence_length and flesch_kincaid_grade.

Moreover, in this study some features such as digits, trigram, colon_count, total_words, punctuation_count, comma_count and semi-colon_count produced negative feature importance scores. A negative importance score signifies that the predictions are less accurate on real data and vice versa on shuffled data. This means that the feature does not contribute much to predictions (importance close to 0), but random chance causes the predictions on shuffled data to be more accurate. For instance, semicolon_count with a more negative importance value has zero contribution to prediction as indicated by a value tending to -0.001. However, predictions on shuffled data will be more accurate because of the effect of random chance on shuffled data.

For purposes of dimensionality reduction and because the document length is a bit longer for document level than sentence level, the feature space can be small but still achieve good results [11]. In this regard we chose to reduce to the feature space and choose features with importance scores greater than 0.002 as the best features for this task. Hence this study reports that the best document level features for separating documents with style changes from documents without style changes were number of sentence repetitions (num_sentence_repetitions), 5-gram and 4-gram.

B. Experiment II

To identify the most determinant sentence level features, all the sentence level features identified from literature were ranked based on their feature importance scores using a random forest classifier. The results of feature ranks and importance scores are shown in table 2.

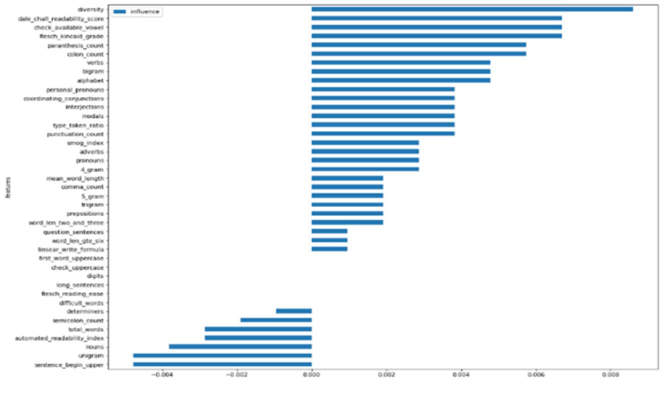


Table 2 : Sentence level feature importance scores.

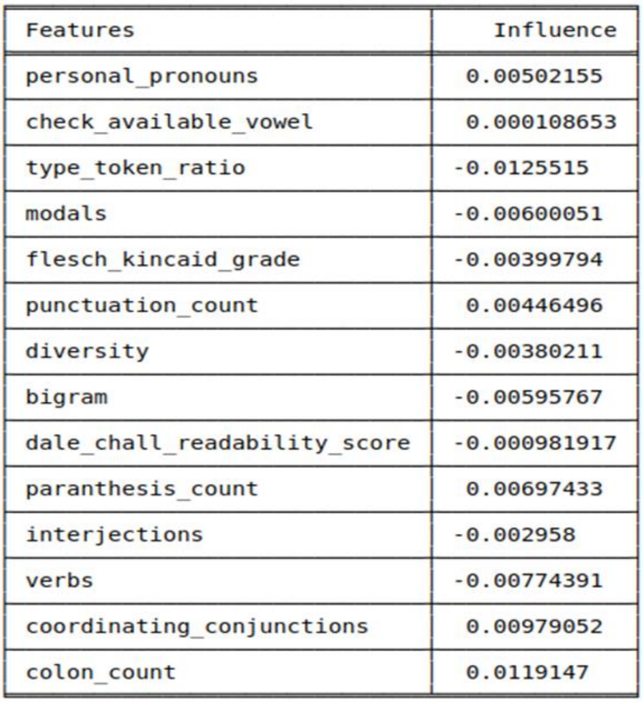
Table 2 presents the results of feature importance for all the features generated at the sentence level computed using Random Forest algorithm. The higher the importance score the more important the feature is. All features with an importance score above zero are considered important while those with zero and negative importance scores are insignificant.

From the above results the most significant sentence level feature for determining the actual number of style changes in a document was diversity at an importance score of 0.009. This was followed by dale_chall_readability score, check_available_vowels and flesch_kincaid_grade at a score of 0.007. parenthesis_count, colon_count all produced a feature importance score of 0.006. verb, bigram, and alphabets also had the same importance score of 0.005. Other top ranked features were at a score of 0.004 were personal_pronouns, coordinating_conjuctions, interjections, modals, type_token_ratio and punctuation_count.

Some features yielded zero importance score indicating that they contribute zero information to model performance. The features were first_word_uppercase, check_uppercase, digits, long_sentences and flesch_reading_ease. Other non-significant sentence level features were difficult_words, determiners, semi-colon_count, total_words, automated_readability_index, nouns, unigram and sentence_begin_upper. These features yielded negative importance scores indicating that they contribute zero information to model performance rather are considered as noise. This study selected features with importance scores of 0.004 and above to be used in the next experiment because they were considered the most determinant. This gave us a total of fifteen (15) features that show importance at the sentence level. Therefore, the study found out that the most significant features for determining the number of style changes at the sentence level were diversity, dale_chall_readability_score, check_available_vowels, flesch_kincaid_grade, parenthesis_count, colon_count, verbs, bigram, alphabet personal_pronouns, coordinating_conjuctions, interjection, modals, type_token_ratio, and punctuation_count.

C. Experiment III

Lastly, we carried out a feature ablation study to determine the effect of each feature on determining the number of style changes in documents. The goal of this step was to find a set of features which yields the best performance measure. The study starte d with all the fifteen features, dropping a feature and measuring the model performance until only one feature is left. This experiment was carried out repeatedly while reshuffling the features. The actual feature contribution, which is the difference between the original performance when a model was trained on the feature and its performance when the feature is removed, was calculated. To evaluate model performance, the study used an ordinal classification index (OCI) measure which measures the error of prediction. Therefore, a lower OCI value indicates a better model performance. The results are presented in fig 3 below.



Features	Influence
personal_pronouns	0.00502155
check_available_vowel	0.000108653
type_token_ratio	-0.0125515
modals	-0.00600051
flesch_kincaid_grade	-0.00399794
punctuation_count	0.00446496
diversity	-0.00380211
bigram	-0.00595767
dale_chall_readability_score	-0.000981917
paranthesis_count	0.00697433
interjections	-0.002958
verbs	-0.00774391
coordinating_conjunctions	0.00979052
colon_count	0.0119147

Fig 3: Feature contributions to model performance

Several runs were done changing the order of which features were first removed and an average of all the run determined. The study presents an average of fifty (50) runs. The influence value is the difference between the original performance when a model was trained on the feature and its performance when the feature is removed. The higher the influence value the more significant a feature is to the model performance. From the results in fig 3 above, some features had positive contributions to model performance indicated by positive influence values while others contribute negatively indicated by negative influence values. Features which yield positive values are considered important while the ones yielding negative influence values have no significance to the model performance.

The removal of each of the following features: type_token ratio, modals, flesch_kincaid_grade, diversity, bigrams, dale_chall_readability score, interjections and verbs resulted in better model performance. This meant that the model performs worse when they are added to the feature indicating negative effect on model performance. These features are not good candidates for determining the number of style changes in multi-authored documents at the sentence level.

Although this experiment used features which ranked top because their higher importance score, the results of this experiment show that some of these features had negative influence on the model performance. Features such as diversity which ranked top as the most determinant feature in experiment two are shown to affect the model performance negatively. Similarly, the actual contribution of dale_chall_readability_score and Flesch_kincaid_grade are negative values indicating their unsuitability for this task.

Other features such as personal pronouns, check_available_vowels, punctuations_count, parenthesis_count, coordinating_conjuctions and colon_count have positive contributions to model performance. Removal of either of the features resulted in a decrease in model performance. In other words, the model performs better when these features are included than when they are removed from the feature set.

Out of the fifteen (15) features used in this experiment, only six (6) features contribute positively to model performance. This shows that there is a 40% overlap in terms of positive importance to this task.

The study selected the all the features with positive influence and combined them to form the best sentence level features. Therefore, this study found out that the best feature combination for determining the number of style changes at the sentence level is the set of features that include personal pronouns, check_available_vowels, punctuations_count, parenthesis_count, coordinating_conjuctions and colon_count

D. DISCUSSION

This study sought to find out the optimal feature set for separating documents with style change from documents without style changes, and to establish the best sentence level features for determining the number of style changes in multi-authored documents. We carried out three experiments.

In the first experiment which ranked document level features the study found out that the most determinant features for the task were number of repeated sentences, 5-grams and 4-grams. These features have produced promising results in previous studies. [21] reported improved model performance when repeated sentences was used as a feature to separate single authored from multi-authored documents. The study observed that there were quite a lot of sentence repetitions in this dataset. We think that this could be part of the reason why this feature ranked highest. Character n-grams such as 5grams and 4-grams have also been effectively applied in style change detection authors and even in authorship verification involving short length text [30], [45]. They have been the regarded as the go to features for authorship studies. Previous studies have used character n-grams with [11] reporting the effectiveness of n-grams upto 5-grams.

We compared the top fifteen features at the document level with the top fifteen sentence level. This comparison was anchored on the fact that experiment ranked all the features including sentence level features. The study observed that some features which ranked top at the document level did not rank well at the sentence level. Features such as first_word_uppercase, sentence_begin_uppercase, difficult_words and automated_readability_index ranked among the top fifteen features at the document level yet yielding zero and negative importance scores at the sentence level. The study opines that the suitability of these features at the document level is anchored on their dependency on the number of sentences. For instance, Automated_readability_index improves with increasing number of sentences and becomes significant when the number of sentences reach thirty and above. similarly, the formula for calculating difficult_words is also dependent on the number of sentences. In addition, sentence_begin_uppercase and first_word_uppercase can only be effective in discriminating between styles if the number of sentences are more than one. The significance of these features greatly reduces if they are extracted at the sentence level.

On the other hand, some features which produced zero and negative importance scores at the document level were observed to yield positive or higher importance scores at the sentence level. Features such as check_available_vowels, flesch_kincaid_grade, punctuations_count and SMOG index produce positive effect at the sentence level. The significance of these features in separating different style decreases with the document length. As the length of the text increases the discriminating attributes might merge and reduce their level of significance.

There were some features which were important at both the document and sentence level. Diversity, dale_chall_readability_score, type_token_ratio, parenthesis_count, verbs, pronouns, prepositions, interjections and bigrams among others. These features are stable style markers for the writing style change detection regardless of the text length.

When the ablation study was performed using the top sentence level features, the study observed that some removing certain features results in decreased model performance. This is despite the fact that these features ranked top based on the feature importance scores. For instance, diversity which ranked as the most determinant sentence level feature yields a negative influence when used in the model. This is because it not a measure the level differences in languages or styles rather, it focusses on the distinctiveness of languages and their frequency as mother tongues. In addition, it may be an effective measure in longitudinal studies involving tracing style changes over time.

III. IV. CONCLUSION

This paper sought to determine the best feature set for the writing style change detection task. To answer this question, we performed exploratory research to identify features which have been applied to the writing style change detection studies. The study engineered features for the two tasks; identifying the most determinant features able to separate documents with or without style changes, and to determine the optimal feature set for determining the exact number of styles in the document. We train a random forest to rank document level features and sentence level features separately, while k-means clustering algorithm is used for the third experiment.

The study concludes that the best document level feature set for separating documents with and without style change was provided by an ensemble of features including number of sentence repetitions (num_sentence_repetitions), 5-grams, 4-grams. These were the top ranked features in experiment one. In addition most previous studies have employed different subsets of these features with success [18], [19], [21]. Consequently, the optimal feature set for determining the number of style changes in documents is provided by an ensemble of features including personal pronouns, check_available_vowels, punctuations_counts, parenthesis count, coordinating conjunctions and colon count. They were among the top ranked features in experiment two. Moreover, they were confirmed as having positive contributions to the model performance using an ablation test.

The study recommends using different feature engineering techniques to further qualify the results of this study. In addition, given the very short length nature of the training set we recommend new character and word level features such as the sentence type be engineered to help with the task.

IV. REFERENCES

[1]. E. Zangerle, M. Mayerl, G. Specht, M. Potthast, and B. Stein, “Overview of the Style Change Detection Task at PAN 2020,” CEUR Workshop Proc., vol. 2696, no. September, pp. 9–12, 2020.

[2]. E. Zangerle, M. Tschuggnall, G. Specht, B. Stein, and M. Potthast, “Overview of the Style Change Detection Task at PAN 2019,” no. September, pp. 9–12, 2019.

[3]. H. Alberts, “Author clustering with the aid of a simple distance measure: Notebook for PAN at CLEF 2017,” CEUR Workshop Proc., vol. 1866, 2017.

[4]. S. Alshamasi and M. Menai, “Ensemble-Based Clustering for Writing Style Change Detection in Multi-Authored Textual Documents,” CEUR Workshop Proc., vol. 3180, pp. 2357–2374, 2022.

[5]. H. Gómez-Adorno, J. P. Posadas-Duran, G. Ríos-Toledo, G. Sidorov, and G. Sierra, “Stylometry-based approach for detecting writing style changes in literary texts,” Comput.

y Sist., vol. 22, no. 1, pp. 47–53, 2018, doi: 10.13053/CyS-22-1-2882.

[6]. D. Castro-Castro, C. Alberto Rodríguez-Losada, and R. Muñoz, "Mixed Style Feature Representation and B 0-maximal Clustering for Style Change Detection Notebook for PAN at CLEF 2020."

[7]. M. L. Brocardo, I. Traore, S. Saad, and I. Woungang, "Authorship verification for short messages using stylometry," 2013 Int. Conf. Comput. Inf. Telecommun. Syst. CITS 2013, 2013, doi: 10.1109/CITS.2013.6705711.

[8]. P. Juola and R. H. Baayen, "A controlled-corpus experiment in authorship identification by cross-entropy," Lit. Linguist. Comput., vol. 20, no. SUPPL. 1, pp. 59–67, 2005, doi: 10.1093/llc/fqi024.

[9]. P. Juola, "Authorship attribution for electronic documents," IFIP Int. Fed. Inf. Process., vol. 222, pp. 119–130, 2006, doi: 10.1007/0-387-36891-4_10.

[10]. H. Ahmed, "The Role of Linguistic Feature Categories in Authorship Verification," in Procedia Computer Science, 2018, vol. 142, pp. 214–221, doi: 10.1016/j.procs.2018.10.478.

[11]. N. Potha and E. Stamatatos, "Intrinsic author verification using topic modeling," Jul. 2018, doi: 10.1145/3200947.3201013.

[12]. M. L. Brocardo, I. Traore, and I. Woungang, "Authorship verification of e-mail and tweet messages applied for continuous authentication," J. Comput. Syst. Sci., vol. 81, no. 8, pp. 1429–1440, Dec. 2015, doi: 10.1016/J.JCSS.2014.12.019.

[13]. A. Abbasi and H. Chen, "Visualizing authorship for identification," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 3975 LNCS, no. April 2016, pp. 60–71, 2006, doi: 10.1007/11760146_6.

[14]. V. A. Oloo, C. Otieno, and L. A. Wanzare, "A Literature Survey on Writing Style Change Detection Based on Machine Learning : State-Of- The -Art- Review," vol. 70, no. 5, pp. 15–32, 2022.

[15]. R. Zheng, J. Li, H. Chen, and Z. Huang, "A framework for authorship identification of online messages: Writing-style features and classification techniques," J. Am. Soc. Inf. Sci. Technol., vol. 57, no. 3, pp. 378–393, Mar. 2006, doi: 10.1002/ASI.20316.

[16]. A. Gelbukh, "Computational Linguistics and Intelligent Text Processing: 16th International Conference, CICLing 2015 Cairo, Egypt, April 14-20, 2015 Proceedings, Part II," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 9042, no. April, 2015, doi: 10.1007/978-3-319-18117-2.

[17]. A. Abbasi and H. Chen, "Applying authorship analysis to extremist-group Web forum messages," IEEE Intell. Syst., vol. 20, no. 5, pp. 67–75, Sep. 2005, doi: 10.1109/MIS.2005.81.

[18]. D. Zlatkova et al., "An ensemble-rich multi-aspect approach for robust style change detection: Notebook for PAN at CLEF-2018," CEUR Workshop Proc., vol. 2125, 2018.

[19]. C. Zuo, Y. Zhao, and R. Banerjee, "Style Change Detection with Feed-forward Neural Networks," no. September, pp. 9–12, 2019.

[20]. Y. Sari, "Neural and Non-neural Approaches to Authorship Attribution," 2018.

[21]. S. Nath, "Style Change Detection by Threshold Based and Window Merge Clustering Methods (Notebook paper) Style Change Detection by Threshold Based and Window Merge Clustering Methods," no. September, 2019.

[22]. W. Daelemans et al., "Overview of the Author Identification Task at PAN 2014."

[23]. S. H. H. Ding, B. C. M. Fung, F. Iqbal, and W. K. Cheung, "Learning Stylometric Representations for Authorship Analysis," Jun. 2016.

[24]. A. Abbasi and H. Chen, "Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace," ACM Trans. Inf. Syst., vol. 26, no. 2, Mar. 2008, doi: 10.1145/1344411.1344413.

[25]. R. Gorman, "Author identification of short texts using dependency treebanks without vocabulary," Digit. Scholarsh. Humanit., vol. 35, no. 4, pp. 812–825, 2020, doi: 10.1093/LLC/FQZ070.

[26]. J. A. Khan, "A model for style change detection at a glance: Notebook for PAN at CLEF 2018," CEUR Workshop Proc., vol. 2125, 2018.

[27]. E. Zangerle, M. Mayerl, G. Specht, M. Potthast, and B. Stein, "Overview of the Style Change Detection Task at PAN 2020," CEUR Workshop Proc., vol. 2696, 2020.

[28]. M. Iqbal, M. M. Abid, M. N. Khalid, and A. Manzoor, "Review of feature selection methods for text classification," Int. J. Adv. Comput. Res., vol. 10, no. 49, pp. 138–152, 2020, doi: 10.19101/ijacr.2020.1048037.

[29]. S. H. H. Ding, B. C. M. Fung, F. Iqbal, and W. K. Cheung, "0 Learning Stylometric Representations for Authorship Analysis," 2015.

[30]. H. Gómez-Adorno, Y. Aleman, D. Vilariño, M. A. Sanchez-Perez, D. Pinto, and G. Sidorov, "Author clustering using hierarchical Clustering analysis: Notebook for PAN at CLEF 2017," CEUR Workshop Proc., vol. 1866, 2017.

[31]. D. Karaś, M. Śpiewak, and P. Sobecki, "OPI-JSA at CLEF 2017: Author clustering and style breach detection: Notebook for PAN at CLEF 2017," CEUR Workshop Proc., vol. 1866, 2017.

[32]. R. Kaur, S. Singh, and H. Kumar, "TB-CoAuth: Text based continuous authentication for detecting compromised accounts in social networks," Appl. Soft Comput. J., vol. 97, Dec. 2020.

[33]. M. Kocher, "UniNE at CLEF 2016: Author Clustering," CEUR Workshop Proc., vol. 1609, pp. 895–902, 2016.

[34]. R. Deibel and D. Löfflad, "Style change detection on real-world data using an LSTM-powered attribution algorithm," CEUR Workshop Proc., vol. 2936, pp. 1899–1909, 2021.

[35]. A. Sittar, H. R. Iqbal, and R. M. A. Nawab, "Author diarization using cluster-distance approach," CEUR Workshop Proc., vol. 1609, pp. 1000–1007, 2016.

[36]. A. Pandian, R. Ragavi, and V. V Ramalingam, "Feature Extraction and Feature Selection process in Authorship Identification for Tamil Language," no. 6, pp. 1–6, 2020, doi: 10.35940/ijrte.F1001.0476S619.

[37]. N. Akiva and M. Koppel, "Identifying distinct components of a multi-author document," in Proceedings - 2012 European Intelligence and Security Informatics Conference, EISIC 2012, 2012, pp. 205–209, doi: 10.1109/EISIC.2012.16.

[38]. S. Nath, "Style change detection using Siamese neural networks," CEUR Workshop Proc., vol. 2936, no. February, pp. 2073–2082, 2021.

[39]. J. Grieve, "Quantitative authorship attribution: An evaluation of techniques," Lit. Linguist. Comput., vol. 22, no. 3, pp. 251–270, 2007, doi: 10.1093/llc/fqm020.

[40]. B. Allison and L. Guthrie, "Authorship attribution of E-Mail: Comparing classifiers over a new corpus for evaluation," Proc. 6th Int. Conf. Lang. Resour. Eval. Lr. 2008, pp. 2179–2183, 2008.

[41]. D. Ghosh, A. Khanam, Y. Han, and S. Muresan, "Coarse-grained argumentation features for scoring persuasive essays," 54th Annu. Meet. Assoc. Comput. Linguist. ACL 2016 - Short Pap., no. Section 2, pp. 549–554, 2016, doi: 10.18653/v1/p16-2089.

[42]. T. Rawat, "Feature Engineering (FE) Tools and Techniques for Better Classification Performance," Int. J. Innov. Eng. Technol., vol. 8, no. 2, 2017, doi: 10.21172/ijiet.82.024.

[43]. M. Koppel and J. Schler, “Authorship verification as a one-class classification problem,” in Proceedings, Twenty-First International Conference on Machine Learning, ICML 2004, 2004, pp. 489–495, doi: 10.1145/1015330.1015448.

[44]. J. S. Cardoso and R. Sousa, “Measuring the performance of ordinal classification,” Int. J. Pattern Recognit. Artif. Intell., vol. 25, no. 8, pp. 1173–1195, 2011, doi: 10.1142/S0218001411009093.

[45]. F. Howedi, M. Mohd, Z. A. Aborawi, and S. A. Jowan, “Authorship Attribution of Short Historical Arabic Texts using Stylometric Features and a KNN Classifier with Limited Training Data,” J. Comput. Sci., vol. 16, no. 10, pp. 1334–1345, 2020, doi: 10.3844/jcssp.2020.1334.1345.

Cite this article as :

Vivian Oloo, Lilian D. Wanzare, Calvins Otieno, "An Optimal Feature Set for Stylometry-based Style Change detection at Document and Sentence Level", International Journal of Scientific Research in Computer Science, Engineering and Information Technology (IJSRCSEIT), ISSN : 2456-3307, Volume 8 Issue 6, pp. 295-313, November-December 2022. Available at doi : https://doi.org/10.32628/CSEIT228617
Journal URL : https://ijsrcseit.com/CSEIT228617