Machines and Algorithms
http://www.knovell.org/mna

Research Article

A Framework for the Authorship Identification in Research Papers

Muhammad Ahmad[1], Muhammad Sanaullah[2, *] and Tanzeela Kousar[3]

[1]Department of Information Technology, Bahauddin Zakariya University, Multan, 60000, Pakistan
[2]Associate Professor, Department of Computer Science, Air University, Islamabad, 44230, Pakistan
[3]Institute of Computer Science and Information Technology, The Women University Multan, 60000, Pakistan
[*]Corresponding Author: Muhammad Sanaullah. Email: muhammad.sanaullah@aumc.edu.pk
Received: 10 August 2024; Revised: 05 September 2024; Accepted: 7 October 2024; Published: 10 October 2024
AID: 003-03-000043

Abstract: Authorship identification and inherent plagiarism detection are crucial to academic and literary ethics. Traditional EPD techniques compare papers to digitalized or internet-available sources, missing plagiarized content from novels or textbooks. Adding non-contributors' names to papers is unethical and undermines motivated researchers' reputation. This study uses stylometric traits to determine authorship and plagiarism without external sources. Stylometric indicators including writing style, language, and sentence structure are used to assign authors to document parts and uncover discrepancies that indicate numerous contributors. Clustering is used to count the authors in a manuscript, solving unethical authorship attributions and concealed plagiarism. The study analyzes methods, identifies limits, and recommends anomaly detection and text feature improvements. The findings show that the suggested method can detect multi-author contributions and non-digital plagiarism. This study provides a complete authorship identification and intrinsic plagiarism detection method to promote academic integrity, discourage unethical activities, and inspire real researchers.

Keywords: Authorship Identification; Intrinsic Plagiarism Detection; Stylometric Features; Clustering Techniques; Academic Integrity;

1. Introduction

A document may be authored by multiple individuals, particularly in the context of research articles, novels, or literature, resulting in recognition and potential financial and career advantages. Within a research community, an individual's reputation is often determined by their publication count; however, this scoring system has led to unethical practices, wherein some individuals coerce their colleagues into including their names on author lists despite lacking any contribution to the research in question. As a result, motivated researchers are experiencing stress and neglect.

To facilitate the motivated researchers by avoiding unethical techniques of publications (author identification/diarization and plagiarism identification) a more comprehensive solution is required, which can detect the contribution of author and plagiarism without requiring the external source from where the text is copied. Such techniques based on the writing styles, words and sentence structuring features of the authors. In research this area can be referred as Authorship Identification or Author Diarization or Intrinsic Plagiarism Detection (IPD).

For authorship identification this work used stylometric features of texts. Stylometric features are used to find the writing style, words and sentence structuring of an author. Stylometric features [3] helps for getting the text features in a document. After getting the text features in this clustering method is used to find number of authors in text document. It encompasses motivation, problem definition, stylometric characteristics, and study scope. Furthermore, the objectives and goals are addressed in the subsequent paragraphs.

Sometimes we face some documents written by number of authors. When we are reading you notice that something seems to be off, style of writing does not seem consistent. However, we can't say that exactly what are those inconsistencies. We need to find that how much authors involved in writing this document and which part is written by whom.

Authorship identification is closely related to text forensic research field of intrinsic plagiarism detection (IPD) and verification of contribution of unknown number of authors in a group assignment in educational field. There can be many examples in real life, the importance of IPD, for example if someone write a text document and copy data from a source that is not digitalized i.e. novels or text book which is not available on internet, then we can't compare it as per our knowledge the document of author with any source. Another aspect is that if one person is doing research and just add name of some other persons and increase their number of publications. So, we can't detect plagiarism in text document and also cannot identify that given article is written by only one person or all mentioned persons are involved in writing this document, by using EPD (external plagiarism detection) techniques. However, EPD cannot detect plagiarism without external source so that unethical activities are being promote and actual researchers who are doing research honestly are disheartened because scammers are increasing their number. Here we need to use authorship identification technique for finding and comparing author's writing style.

Since the field of plagiarism is very vast, there is a lot of digital text now a days available in form of blogs, digital novels, and scientific papers etc. The main field of plagiarism is academic. In academic researchers write their articles or research papers are published and researchers add names of other persons as contributor but actually they have no contribution in that research or writing that article. On the other hand, students need to write their thesis, a scientific paper written by a student or a group of students. So, we need to identify how many authors involve writing a thesis or paper. This can be possible through Authorship Identification. We need to extend the Authorship Identification techniques. So, we can get better results. We can do this by detecting more and valuable text features and by applying the anomaly detection techniques. PAN 19 focused on two tasks one finding that in document multi authors involve or not and second for finding total number actual authors.

Our aim in this research work is to answer the following questions:

- What current work in the field of authorship identification using technique of clustering has done previously?
- Which stylometric features have been used for authorship identification by other researchers?
- Which stylometric features we should use to improve our results and why?
- What are the advantages of stylometric features used in our approach?

Rest of the paper is organized as follows. In Section 2 we will provide overview of the existing literature of Authorship identification tasks. It also explains the available methods for Authorship identification techniques. In Section 3 the proposed approach for Authorship identification task we used will be explained. In Section 4 the results obtained from our proposed approach will be discussed and Section 5 will conclude the paper.

2. Literature Review

This section is bifurcated into two parts. The initial section will address the Pan Plagiarism Competition (PAN-PC) about authorship identification or intrinsic plagiarism. The second chapter will address the study on authorship identification conducted by scholars..

2.1. PAN Plagiarism Competition (PAN-PC)

Plagiarism detection is a critical issue, particularly in the realms of academia and research. The most significant aspect of plagiarism is its automatic detection. A lot of software is available in market and different researchers have worked in this field and published their papers. Various algorithms are proposed by different researchers and a lot of algorithms are available on internet, but it is very difficult to guess that which algorithm is best for plagiarism detection. This problem was overcome by PAN-09 [1], they hold a competition.

2.1.1. PAN-09

For the first time an initiative taken by PAN was to organize a competition on plagiarism detection. They setup a controlled evaluation environment for plagiarism detection. They managed a controlled evaluation environment which contain quality measures of measure and corpus which have large plagiarism. Future plagiarism detection research could be compared by unified test environment which they provided. They set up a corpus consisting of large-scale plagiarism *(Dq, D, S)*, where source documents collection called as D, suspicious documents collection called as *Dq* and set of annotations of all plagiarism cases between *Dq* and *D* called as *S*. They divided the competition into two phases. Different symbols were not denoted the sub-corpora.

1. **External Plagiarism Detection Task:** In this task given is D and Dq the task was to identify sections in D which are source sections and Dq which are plagiarized.
2. **Intrinsic Plagiarism Detection Task:** In this task given is Dq In IPD task Plagiarized sections needed to identify without given any sources. In their system there was a corpus consisting of large scale for artificial plagiarism and detection quality measures. In Pan-09 they provided 41,223 text documents and in which they provided 94,202 cases of artificial plagiarism.

2.1.2. PAN-10

PAN-10 [13] was an enhanced version of PAN-09. In this competition the corpus was made to assess the system's execution. It had both manual and programmed plagiarized instances. This corpus contained 68,558 plagiarized text documents. The improvement in the evaluation framework was the main agenda of this competition, because in every research field this is a serious problem. They also introduced detection granularity, that is used to recognize the in plagiarized text passages. Low granularity efficient the review of algorithmic identified sections and style of an algorithmic examination inside a process. They applied three measures combined but these three can be isolated as signal for overall performance score.

2.1.3. PAN-13

In PAN-13 [2] the author identification task focused verification of authors in documents, documents are provided as a set of a questioned document and a single author, the task was to identify in the set of documents the particular was involved to write the questioned document or not. As well as In the competition they presented performance measures, the new corpus, the evaluation setup they built for task. They were covering three different languages for this task.

Performance Measures

In PAN-13 participants provided answers of each problem in simple binary "yes/no" for the author identification task. In case if their provided solution not able to answer some problem then leave unanswered them. To evaluate them PAN-13 used the following measures:

$$Precision = \#correct\_answers / \#answers \quad (1)$$

$$Recall = \#correct\_answers / \#problems \quad (2)$$

This showing that if they answered all the problems then Recall and Precision measures are equal. So, they computed ranking for the whole evaluation corpus of all languages by combining above mentioned measures via F1.

2.1.4. PAN-14

The corpus was comprising with four natural languages (Spanish, Greek, English, and Dutch) and also from different four genres (novels, reviews, essays, opinion articles). In addition, in this competition the focus on the accuracy, more suitable performance measures and the confidence of the predictions were used.

2.1.5. PAN-15

Authorship, Social Software Misuse and Uncovering Plagiarism focuses on that direction a series of evaluation labs. In PAN-15 [3] edition they were comprised 3 problems

1. ***Plagiarism Detection:*** In this problem the task was to detection of plagiarized sources and also re-used passages' boundaries in a given document.
2. **Author Profiling:** In this problem the task was to extract information about the author in a given document like age, gender etc.
3. **Author Identification:** Identify its author in a given document.

2.1.6. PAN-16

In Pan-16 [4] competition they divide intrinsic plagiarism task into three sub tasks. First task related to traditional intrinsic plagiarism task in which need to identify the text in document related to which author (main author or others). In second task the number of authors given and need to identify which text of document related to which author. In third task there is unknown number of authors and need to identify how many authors contribute to write a document and which text in document related to which author.

1. ***Tasks and Corpora:*** In Pan-16 the shared task focused on identification of authorships in a single document. They chose a title Author Diarization for all of its three related sub problems.
2. ***Traditional Intrinsic Plagiarism Detection:*** In Traditional Intrinsic plagiarism detection assumed as a document written by an author. That writer involved in written of document at least 70%, The problem is to identify that the remaining text portions written by others.
3. ***Diarization with provided no of authors:*** In this problem they were given the exact number of authors that were involved in written a document, the problem is to find the contribution of each author in a given document.
4. ***Unrestricted Diarization:*** In this problem the number of authors not given we need to identify how many authors involved to write a given document and also which portion of texts in a document written by which author is called unrestricted diarization.

2.1.7. PAN-17

Style breach detection a document is given to determine in this document multi-authors are involve or not and if yes then find the boarder that where author switch**.** It is very difficult to find the task that where is the exact character position where author switched. None of competitor performed better than slightly change in random baseline. [5]

2.1.8. PAN-18

In PAN-17 problem didn't solve accurately, in PAN-18 committee who organize this competition relaxed problem for the competitors in edition 2018.

Style change detection a document is given to competitors to decide whether this document involve one author or more than one authors in document. This task was solved accurately by researchers and problem was solved with high accuracy of 0.89. [6]

2.1.9. PAN-19

As PAN-18 was solved successfully, PAN-19 was built on the base of success of PAN-18 and task divided into two connected sub tasks.

1. 1st task includes to find whether document is written by one or more than one author, i.e. style change exists or not?
2. 2nd was to find that if a document consists on more than one authors then how many original authors are involved in writing this document.

Note that first task is from PAN-18 which is already solved with high accuracy in PAN 18 competition. [7] The second task is simplified form from PAN-16 3rd task, which only requires to find number or authors but do not require to find the same portion of the text.

2.2. Authorship Identification

Since research in authorship identification has been increased in last decade. Because in external plagiarism detection there is need of external source of text to compare to check plagiarism, an external source will always need a source text to compare but in authorship identification/ author Diarization there is no need for external source of text to compare to check plagiarism. Comparison actually based on the finding style change anomaly detection. To find anomalies there is need to divide text in to fragment of text which are separated by sentence length and passages depend on depend on researcher. Then some attributes would be found out by applying some stylometric features and find distance of each fragment.

2.3. Stylometric Features

In authorship identification we need to identify writing style of authors. Stylometric features are used to detect different writing styles. These features are used to quantify aspects of different writing styles. Here we use an example that some authors use word 'The' again and again, but some authors do not use it repeatedly. Each writer has his own writing style and thinks in his own way. These word frequencies will differentiate one authors style from another. One more thing which we can discuss about authors style is that some authors use long sentence and some use small sentences, this deviation or writing style can be detected by using lexical features. After detecting writing style of text, we need to detect anomaly that how much authors are involve in this text. For this we use anomaly detection technique. Stylometric features are categorized into following features according to Efstathios Stamatatos et all. [3]

- Lexical
- Character
- Syntactic
- Semantic
- Application Specific Features

2.3.1. Lexical Features

Lexical diversity, sentence duration, word length, etc. Word frequencies, n-gram frequencies. As illustrated in Table 1.

Table 1: Lexical Features

List of Lexical Features
Sentence and Word length, etc.
Frequencies of Words
Word length
Frequencies of Word n-grams

Sentence length is to count total number of words used a sentence, word length is total count of characters in a word, frequencies of words are counted to compare that is how much word are being used with high frequency and low frequency. High frequency words are those which are used vey commonly like word 'the'. Frequency n-gram is the summed or mean frequency of all fragments of a word given length.

2.3.2. Character Features

Character types such as digit count, letter count, uppercase letter count, etc., and n-gram analysis. Variable-length character; compression techniques are presented in Table 2.

Table 2: Character based Features

List of Character based Features
n-grams Character having fixed length
Character n-grams having variable length
Character types i.e., digits, letters, etc.

An effective method for representing text for stylometric analysis is the use of n-grams, which may discern subtleties in stylometric characteristics. A technique utilizing variable-length n-grams is employed for online writing. The type of character refers to the overall count of numbers or letters utilized in the text.

2.3.3. Semantic Features

Synonyms, Functional, Semantic dependencies shown in Table 3.

Table 3: Semantic Features

List of Semantic Features
Synonyms
Semantic dependencies parsing
Functional

Synonyms are the words which have same meaning semantically and being used in a certain text. For example, little or small. SPD is task of mapping sentences into a formal representation, in the form of directed graph, of its meaning with the curves between words. Functional words are used to express relation of words with other words grammatical and structural relation.

2.3.4. Application Specific Features

Some characteristics are structural, some merely specific for content, some special for languages that users can utilize displayed in Table 4.

Table 4: Features Specific to Application

List of Application Specific Features
Structural Features
Features specific for Language
Features specific for Content

2.3.5. Syntactic Features

Part-of-Speech, abbreviated as POS, encompasses phrase and sentence structure, frequencies of rewrite rules, and errors presented in Table 5.

Table 5: Syntactic Characteristics

List of Syntactic Features
Frequencies of Rewrite rule
Phrase and Sentence based Structures
Parts of Speech

The pattern matching method is employed to determine the frequencies of rewrite rules. A phrase is a collection of words that cannot stand alone as it lacks both a subject and a predicate. POS tags are referred to as grammatical tags. Parts of speech are utilized in POS tags. Part-of-speech (POS) tags serve as features in a text and can be quantified by the total count of any specific part of speech utilized in the text.

2.4. Intrinsic Plagiarism Detection

The intrinsic plagiarism detection problem, as defined in section 1.2, aims to identify the optimal features for detecting stylistic changes in text documents authored by multiple individuals, where variations in writing style occur within the content. In the following sections, we examine the pertinent literature on authorship identification.

Stamatatos et al. [8] address a conventional intrinsic plagiarism problem. In their study, a document is segmented using a sliding window approach [3], and character n-gram profiles are employed to discern authorial styles in the text. Their method involves automatic segmentation of documents based on stylistic variations to determine the presence of plagiarism. Their methodology involved defining a sliding window over the text length, within which they compared the text to the entire document. The anomalies were utilized to identify the plagiarized passages. Subsequently, the entire document identified probable plagiarized segments that exhibited significant dissimilarity from the relevant text sections.

Chaoyuan Zuo et al. [11] segmented the material into multiple parts and clustered them based on writing style. Their primary objective was to ensure that the overall number of clusters matched the total number of authors in the submitted document. They utilized documentation in both Spanish and English. However, hardly 1% of the documents in their collection were in Spanish. The documents randomly assigned a total of 1 to 5 authors. Subsequently, they eliminated several common phrases that possessed minimal or no grammatical significance. Subsequently, conducted binary categorization of publications including one or several authors. A category was created for many writers and single authors, utilizing Keras for this implementation. To ascertain the number of authors, the document is segmented and clustered. Several texts were inadequately organized, and Chaoyan Zuo et al. employed the NLTK tokenizer; some documents produced over 200 sentences, with the documents segmented at the paragraph level rather than the sentence level. It was determined that style changes were identified at the beginning of a new line or following an empty line. 80% was noted subsequent to the newline.

Sukanya Nath [12] employed a strategy to segment a text material into paragraphs. The window was to be regarded as equivalent to the paragraph. The window tokenizer was adjusted to combine extremely small paragraphs, specifically those under 200 characters, with the previously studied paragraph. The lengthy paragraphs were divided into smaller sections to achieve a balanced window length. A method called window merge clustering was employed to amalgamate all analogous windows, resulting in a new set of windows. Utilizing these new windows, they computed the distance matrix for the subsequent iteration. This procedure resulted in the formation of hierarchical clusters. The objective was to depict each cluster as a collective depiction of its constituents, rather than focusing on individual distances.

Elamine et al. [13] concentrated on stylometric characteristics that most effectively delineate writing style, employing hybrid elements. They recommended a five-step procedure. Initially, they categorized the documents based on writing style. In the second step, they tokenize each obtained cluster into segments of 500 characters to facilitate feature utilization in subsequent rounds. In the subsequent stage, they generated vector characteristics and developed a style function to ascertain the style for each cluster. The third phase was detecting outliers.

Akiva [9] also addressed the issue of intrinsic plagiarism detection, which was the primary focus of the PAN-11 competition. The author employed a methodology comprising two phases: chunk clustering and chunk property detection. Initially, they partitioned the provided document into segments of 1000 characters. The author identifies the 100 most uncommon words utilized in at least 5% of the pieces. The author subsequently generated a numerical vector representing segments, with a length of 100, to identify the presence or absence of rare terms inside those fragments. The cosine metric was employed to assess similarity between pairs of pieces. The spectral clustering method, commonly known as n-cut, was employed for clustering the segments. The document was categorized into two sections by the author: plagiarized text and original text. The objective of the subsequent phase was to identify the plagiarized sections inside the document. The author employed a clustering technique on the training corpus and assessed many attributes, including the absolute and relative sizes of all clusters, the similarity of each segment to the entire document, to other clusters, and to its own cluster. The author disregarded any documents exhibiting over 40% plagiarism and thereafter picked random excerpts from the remaining texts. The author attained a recall of 6.6% and a precision of 12.7% in the assessment of PAN-11.

Oberreuter and Velásquez [10] investigated intrinsic plagiarism detection by the analysis of variations in writing style. Initially, the documents underwent pre-processing, wherein all characters were eliminated, retaining just those inside the a-z range, and all characters were converted to lowercase. Subsequently, they examined word unigrams, taking into account all terms, including stop words. Subsequently, word-frequency-based algorithms were employed to assess the similarity of the manuscript. A frequency vector was constructed for all words, and subsequently, the papers were clustered into groups. The author initially produced these pieces from the entire documents using a sliding window of length 'm'. A new frequency vector is calculated for each segment, which is subsequently analyzed in following phases. This vector is utilized to ascertain deviations from the whole document section. All segments are grouped according to their distance and document style. The author's methodology was assessed using PAN corpora. Standard measures were employed to assess their approach to information retrieval. The results derived from their methodology exhibited an untrustworthy nature due to an exceedingly low precision of 0.3.

In Pan-16, Sittar et al. [14] engaged in the author diarization task, employing varying quantities of text to segment documents and utilizing lexical and character features to identify authors' writing styles. For Task A, they utilized sentence counts of 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, and 15; for Task B, they employed sentence counts of 5, 10, 12, 13, and 14; and for Task C, they used sentence counts of 5, 10, and 12. Table 2.5 presents the sentence lengths from Sittar et al. [14]. They employed clustDist [15], a simple method to ascertain the average distance from one segment of text to all other segments, then calculating the mean of all resultant distances. Consider a document D containing n sentences, where each phrase i is identified by calculating p features to generate a feature vector Vi for that sentence. A matrix V of dimensions n*p was constructed for their research, with each row representing a feature vector of a text.

ClustDist is calculated using equation (3), where d is the distance between any two vectors.

$$ClustDist(a, B) = \frac{\sum_k d(a,b)}{n} \quad (3)$$

The resulting score for each sentence's distance from others gives a ranking that indicates how a sentence differs from all other sentences in the document.

Kuznetsov et al. [16] employed a sliding window approach [3] to partition a document into fragments. They employed n-gram frequencies, word frequencies, and parts of speech tags as text features, comparing and analyzing them for author diarization. Their proposed solution utilizes a per-sentence approach [17] for

segment creation. In contrast to the sliding-window approach, the sentence method builds discontinuous parts of varying lengths to do sentence-level plagiarism detection. They utilized the standard nltk parser, namely sent_tokenize from the Natural Language Processing Toolkit, to segment the document into sentences.

Polydouri et al. [19] also addressed the issue of intrinsic plagiarism detection. The author employed the sliding window technique for text segmentation. They established a window size of 15 sentences and a window step of 5 sentences. The author employed 11 features for style analysis, encompassing both stylistic and semantic elements. The authors developed a straightforward technique that aims to illustrate potential distribution through compression rate.

Kuznetsov et al. [16] also addressed the issue of intrinsic plagiarism detection. The authors initially partitioned the material into smaller portions. The author addressed the issue of author diarization by modifying the technique of intrinsic plagiarism. An algorithm was employed to segment the document into sentences, which were subsequently vectorized. A train model is employed by the algorithm, and a series of statistics is produced as $a(s_1)$, …, $a(s_m)$, while the sentences are represented as s_1, …, s_m. Concealed The diarization method employs a Markov Model approach with Gaussian emissions to deliver a segmentation series $a(s_1)$, …, $a(s_m)$.

To address an indeterminate number of writers, the authors implemented a method including the computation of an estimated average t-statistic across the segments of all authors. The subsequent equation is employed.

$$Q(n) = \sum_{i,j=1}^{n} \frac{|m(c_i) - m(c_j)|}{\sqrt{\frac{\sigma(c_i)2}{i(c_i)} + \frac{\sigma(c_i)2}{i(c_j)}}} \quad (4)$$

Q(n) = the measure of clusters discrepancy

$m(c_i)$ = mean of elements in cluster

$\sigma(c_i)$ = mean deviation

$I(c_i)$ = cluster size

Bensalem et al. [25] also addressed the issue of intrinsic plagiarism detection. The author initially segmented the text document into multiple parts. Subsequently, these segments are characterized by specific properties. Authors employed a classification algorithm utilizing certain features to train the dataset. These phases facilitate the execution of the author's methodology. Segment the provided document d into fragments s_i using the sliding window approach. S represents the quantity of fragments. The author constructs a model of n-gram documents, excluding numerals. The frequency of n-gram ng_i is utilized to assess its occurrence within document d. If ng_i appears alone once in document d, then its frequency is 1. The highest value can match the whole number of pieces when ng_i is present in each fragment, $s_i \in S$. A vector of m features fi is utilized to represent each fragment s. Fragment vectors derived from all corpus documents are consolidated into a single dataset by the author. All vectors were labeled with authenticity, indicating whether they were plagiarized over 50% or original. The classification process was executed using the WEKA tool.

Tschuggnall et al. [5] also addressed the issues of intrinsic plagiarism detection and stylistic violation detection. Authors employed classification techniques for the identification of style breaches. The performance of the submitted algorithms was evaluated using two criteria commonly employed in the field of text segmentation. The windowdiff metric was proposed for evaluating text segmentation, and it remains applicable to similar issues. The error rate, determined by windowdiff, ranges from 0 to 1, where 0 signifies flawless prediction of borders. Authors utilized various types and datasets according to the challenge, employing a text segmentation approach to report windowdiff values, with 0.01 considered almost perfect and values exceeding 0.6 reported under specific conditions. The WinPR metric is a contemporary

implementation of windowdiff, wherein the author employed this methodology to compute precision and recall through information retrieval using windowdiff. The computation of true and false values was employed to determine WinP and WinR. The evaluator script employed tokenization to calculate these two measures based on character position.

Liu et al. [26] addressed the issue of style crack rearrangement. The authors partitioned the manuscript into segments of text. They employed a range of characteristics to identify style crack. Utilized features include lexical elements, specialized punctuation, synonyms, and functional terms. Authors previously conducted segmentation on materials prior to authorship identification. The authors aimed to identify the crack point by these segmentations. The sliding window technique is employed. Each slide window consists of five sentences simultaneously. When a change in style happens, both the current style and the previous style will ultimately converge until they are identical. The presence of five sentences with minimal information increases the likelihood of accidental occurrences. Authors assert that style changes occur at the conclusion of a paragraph. It was presumed that each paragraph is authored by a singular writer, with stylistic discrepancies manifesting at the conclusion of one paragraph and the commencement of the subsequent one. The sole method to enhance accuracy was to diminish recall. The weights of all criteria were required to investigate style cracks. Adjusting the weights may reveal the style crack.

In the subsequent phase, authors aggregated styles. The primary technique for feature extraction employed is clustering, so the authors utilized style clustering. A mapping association was established between the features and the article. The input for the final k-means was derived using feature extraction. A newspaper corpus was utilized, and 1,300 items were chosen. Of the 1300 articles, 150 were designated as a test set, while 1150 items were utilized as a training set. Twenty news stories were picked from five authors for the experiment. The articles were divided by paragraph. The sliding window technique was employed for each sentence. The clustering findings were suboptimal due to the sentence-based approach, as the variations were minimal. The clustering results were ambiguous. The authors eliminated the sliding window approach. Authors employed a methodology that treats each paragraph as an individual author. Style feature extraction was conducted on each paragraph, followed by the use of the k-means algorithm. The application of this strategy enhanced the results. The paragraph-based approach is superior to the sliding window method for crack pattern recognition.

Seaward and Matwin [27] employed a complexity metric for plagiarism detection in textual documents. Kolmogorov Complexity Measures serve as a stylistic trait for identifying inherent plagiarism. Text segments were generated according to word class, encompassing nouns and non-nouns. The authors utilize the following equation to quantify complexity.

$$K_c(x) = \frac{Length(C(x))}{length(x)} + q \tag{5}$$

- K(x) = Kolmogorov Complexity
- C = compression algorithm

The authors employed two classifiers, Support Vector Machine (SVM) and Neural Network (NN), for training and testing purposes. Precision and recall results were computed on a per-chunk basis rather than for individual characters.

Safin and Ogaltsov [28] addressed the issue of intrinsic plagiarism detection by the application of text statistics. The corpus was derived from the Stack Exchange network, comprising users' posts. The authors initially partitioned the data into a test set and a training set. The authors employed accuracy score to evaluate the quality of the suggested method. The accuracy of binary classification is defined by the authors below.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{6}$$

Where,

TP = True positive

TN = True negative

FP = False positive

FN = False negative

The model employed by the authors has three independent classifiers. Authors utilize Statistical, Counting Classifier, and Hashing classifiers. These classifiers yield probabilities for textual content that reflects stylistic alterations. Final results of probability can be calculated by using weighted sum of p_s, p_h, p_c respectively.

For the final accuracy score weighted sum of probabilities is calculated in text d.

$$\text{Score(d)} = a_s p_s + a_h p_h + a_c p_c \tag{7}$$

To maximize the accuracy, Coefficients and threshold were tuned by using a validation set. The importance of matching classifiers was shown by each coefficient. For the final model optimal parameters are

$$a_s = 0.4,\ a_h = 0.2, a_c = 0.4 \tag{8}$$

As here are the most informative classifiers being statistical and counting classifiers. And the value of δ is 0.55 relation between the accuracy score and threshold value.

Grubisi and Pavlovi [29] addressed the issue of author diarization in PAN-16. The approach was employed to breakdown text materials, creating segments, with each section attributed to an author. The authors suggested a technique that delineates a pipeline comprising three transformations: feature extraction, feature transformation, and clustering. f_b denotes the feature extractor, f_t represents feature transformation, and f_c signifies clustering. If a document D has n tokens, a sequence of n n_b-D features representing the tokens is the output of the feature extractor. At the conclusion of the pipeline, clustering was performed on vectors to identify stylistic elements from the text utilizing the feature vector. The authors employed clustering as the concluding step. Clustering was performed using feature vectors that represent stylistic elements. The total number of clusters obtained corresponds to the number of authors.

Recent studies have brought important innovations in the use of machine learning and natural language for authorship identification. For example, author in [31] presented a new approach to incorporating BERT embeddings and stylometric metrics and outperforming the others for authorship identification on reference datasets. Author in [32] explore the use of deep learning, more specifically neural networks, for stylometric analysis. In order to avoid conventional manual feature extraction, the authors use convolutional and recurrent neural networks to process all text features.

As discussed, that literature work shows no standard format for plagiarism and authorship identification. There is a need to find in a document which portion of a document written by which author or how many authors involved to write a document this is called Authorship identification. In the following section we discuss the method of authorship identification considering the style from text by using stylometric features.

3. Proposed Approach

We need to extract the styles of authors from a document using stylometric features and using an anomaly detection technique for find the distance between the text features from others in a document. The stylometric features for extracting the features of texts is the technique that we used in our proposed approach and then discuss the clustering technique for finding the difference between text features.

3.1. Stylometric Features

While writing a document authors left behind some personal traits in texts unintentionally, that show everyone has his own format for writing a document, therefore we can distinguish the authors from a document by getting the style. For getting the styles of authors we need to identify the features of text from

a document, these features are called stylometric features. These features used for detect the writing style of authors as we have discussed in section 2.

In extracting syntactic features, the analysis is performed on the original text including the stop words to determine features which depend on the stop words such as the percentage of the number of pronouns used, the determiners and conjunctions used in the text. Once these features are extracted, stop words are nothing but the most Frequently used words in natural Language Processing, which are removed and then other features like syllable based long/short ratio, and lexical richness is calculated on the specified text after filtering the stop words.

Table 6: Stylometric Features Proposed by Authors

Author Name	Lexical features	Semantic features	Character features	Application Specific Features	Syntactic Features	Readability Features	Vocabulary Richness Features
Zuo et all				✓	✓	✓	
Sittar et all	✓						
Kuznetsov et al.	✓				✓		
Polydouri et all	✓	✓					
Seaward and Matwin	✓				✓		

The features we used in our approach are following:

3.1.1. Lexical Features

We used following lexical features for identifying the writing styles from text Average Word Length, Average Sentence Length by Word, Average Sentence Length by Special Character Count, Average Syllable per Word, Functional Words Count, Punctuation Count. These used features are very basic features which can be extracted from text. Structure of the text can be known by these features. For example, averages of different counts can be calculated like functional words, Punctuations, word lengths, special characters. Functional words can be used for expressing all grammatical relationships of all words within a sentence. Second thing is that, a word can be most likely a difficult word if it has more syllables (not necessary). The measure of complexity of a word is being average syllables per word, which is used to calculate many other features which are related to readability score. Different genres can be differentiated by using straight way of special character count and punctuation count.

3.1.2. Character Features

We used following character features for identifying the writing styles from text ratio uppercase letters count, character counts, words count, letters count, ratio of spaces, ratio of letter, ratio of tabs, tabs count.

3.1.3. Vocabulary Richness Features

Many contemporary quantitative research increasingly depend on the concept of word richness. We utilize vocabulary richness attributes to discern writing styles from text. The writing style of two authors can be distinguished when a document exhibits low vocabulary richness, characterized by repetitive word

usage and limited lexical variety, whereas a document has high vocabulary richness if the writer employs diverse and novel language. Utilizing these qualities allows us to distinguish between the writing styles of two writers, providing insights into the diversity and language richness present in their texts. Our technique utilizes Hapax Lego Menon, Hapax DisLegemena, Honores R Measure, Sichel's Measure, Brunet's Measure W, Yule's Characteristic K, Shannon Entropy, and Simpson's Index.

1. Hapax Lego Mena and Hapax DisLegemena

A hapax legomenon is a term that appears just once inside a certain context, whether in a single text or across the written corpus of an entire language. This term is occasionally misapplied to denote a word that appears multiple times inside a specific work by an author. Hapax legomenon is a Greek term meaning "(Occasionally) articulated (only) once." Similar to Hapax, DisLegemena is a term that appears twice. The remaining aspects are now elucidated. The subsequent concepts will be employed for their elucidation.

1. Tokens N length in words of text.
2. Count of distinct words type V in text.
3. Count of unique words in the text just once, V1 Hapax Legomena.
4. Count of terms appearing in text exactly twice, V2 DisLegemena.
5. Count of occurring words i times, Vi.

The type/token ratio is influenced by text length; yet, it is a valuable metric for assessing vocabulary richness when comparing texts of identical length.

2. Honore's metric (R)

It is dependent on the hapax Legomena [20]:

$$R = 100 * \log N / (1 - (V1 / V)) \tag{9}$$

3. Sichel's metric (S)

It is dependent on the DisLegemena, and with respect to N it is relatively constant [21]:

$$S = V2 / V \tag{10}$$

4. Brunet's metric (W)

The equation for this measure is mentioned below:

$$w = N^{v-a} \tag{11}$$

where a is a constant (usually 0.17). To be relatively W was found unaffected by text length and to be author specific [22].

5. Yule's characteristic (K)

It is dependent on words of all frequencies [23]:

$$K = 10{,}000 * (M - N) / (N*N) \tag{12}$$

6. Shannon Entropy

Typically, a system's calamity can be induced by entropy. This concept is employed in our text project. Claude Shannon is the progenitor of information theory. He provided Shannon's entropy formula to quantify the information of a certain word.

$$E = \sum_{i=0}^{N-1} P_i log P_i \tag{13}$$

P shows the probability of words occurring in the text [16].

7. Simpson's index

The assessment of diversity can be conducted using Simpson's diversity index. Biodiversity of habitats is frequently quantified using Simpson's diversity index. It considers the prevalence of each species

alongside the current species count. Simpson's Index (D) quantifies the probability that two randomly picked individuals from a sample will belong to the same species. This idea is employed in natural language processing to identify the diversity of text segments. To identify diversity across various parts of text, we employed biodiversity in our project.

$$S\ Index\ (D) = \sum(n/N^2) \tag{14}$$

N = total number of words in a text.

n = total number of unique tokens

3.1.4. Readability Score

A reader can easily understand a document of readability is easy. Readability is a measure of how easy a reader can understand written document and even a letter or character. Researchers are using frequently readability features in the field of linguistics and linguistic 'laws' to use these readability features to calculate readability scores in text. Some features we are using for readability scores are Flesch Reading Ease, Flesch-Kincaid Grade Level, Gunning Fog Index and Dale Chall Readability Formula.

1. Flesch Reading Ease

In 1948 Flesch reading ease was created as a test of readability [21]. This test tells us that how much education is needed to read a piece of document text easily; this test scores tell us roughly. Between 1-100 scores are generated by reading ease formula. To interpret scores a conversion is used. For example, is readability score is generated between 70-80 then it is equal to school grade level 7. It should be easy for and average reader to read a text which have readability score of 70-80. By doing research in education sector Flesch reading ease test originated.

$$FR\ Score = 206.835 - 1.015\left(\frac{total\ words}{total\ sentences}\right) - 84.6\left(\frac{total\ syllables}{total\ words}\right) \tag{15}$$

2. Gunning Fog index

In linguistic, for English writing readability test is Gunning Fog Index. To understand text document on first reading, the index estimate, how much education is needed to a person. Reading level of high school senior of U.S is required if Fog index is 12. Gunning fog index can be calculated by using given formula.

$$G = 0.4 * \left[\left(\frac{words}{sentences}\right) + 100\left(\frac{complex\ syllables}{words}\right)\right] \tag{16}$$

Words consisting three or more syllables are 'complex'.

3.1.5. Syntactic Features

We used following syntactic features for identifying the writing styles from text percentage of nouns, average syllable per word, percentage of words with one syllable, percentage of words with more than three syllable, percentage of pronouns, percentage of personal pronoun, percentage of modal, percentage of verbs, percentage of adjectives, percentage of adverbs, percentage of coordinating conjunction, percentage of subordinating conjunction, percentage of interjections, percentage of determiners.

List of features used in our approach are shown in table 7.

Table 7: Stylometric Features Proposed in Our Approach

Lexical features	Semantic features	Character features	Application Specific Features	Syntactic Features	Readability Features	Vocabulary Richness Features
✓		✓		✓	✓	✓

Name of features used in our approach are shown in table 8.

Table 8: Name of Features Used in Our Approach

Feature Type	Feature Name
Lexical Features	Mean Lexical Length
	Punctuation Frequency
	Functional Words Frequency
	Mean Syllables per Lexeme
	Count of Special Characters
	Mean Sentence Length by Character
	Mean Sentence Length by Character
Character Features	Characters Frequency
	uppercase letters Frequency
	Spaces Frequency
	Tabs Frequency
	Lexical Frequency
	Ratio of Uppercase Letters
	Digits Frequency
Vocabulary Richness Features	Hapax Legomenon
	Shannon Entropy
	Simpson's Index
	Brunets Measure
	Yules Characteristic
	Honores Measure
	Sichel's Measure
	Hapax DisLegemena
Readability Features	Flesch Reading Ease
	Dale Chall Readability Formula
	Gunning Fog Index
	Flesch-Kincaid Grade Level
Syntactic Features	Nouns count
	Verb count
	Adjective count
	Adverbs count
	Pronouns count

3.2. Dataset Selection

We selected our data set 'corpus-webis-trc-12', which encompasses about 150 different topics written by number of authors, same topic written by different number of authors, different authors to different

difficulty level. These were written by professional writers from different places. When, we want to cluster different writing styles, we use 'corpus-webis-trc-12' dataset to perform clustering. The main purpose of this dataset is to demonstrate our approach, but our approach can be used on any kind of document.

3.3. Data Pre-processing

After selecting dataset which consist on about 150 topics and each topic is written by more than 20,000 different authors and each author has his own writing style. First of all, we took random writing styles from each topic which vary from 1 to 5 writing styles and paste them in a single text file arranging text files according to their topic. For example, from topic number 100 we took 5 different author styles and paste them in a single text file named topic100_5, 100 is topic number and 5 is total author styles in this file. In next phase a document is divided into chunks. We set up the size of each chunk equal to 10 sentences because if chunk size would be too large then it was difficult for us to extract the crux for each passage and if it would be too small then it might lose its significance. That's why we used an average of 10 sentences for each chunk and we also can change size of chunks according to need. After dividing into chunks first of we compute lexical features for each chunk of text. For the rest of all features punctuations and special characters are removed and tokenization is performed. Before removing punctuation and special characters we performed lexical features because punctuation and special character are used to perform lexical features.

The choice of a chunk size equal to 10 sentences was informed by balancing two critical factors: In other words, meaningful context retention and computational efficiency. The problem of large chunks is that it becomes hard to achieve feature specificity of the particular writer, as too much text harms the distinctiveness of the central topic by adding noise from other related areas. On the other hand, the sizes which are too small can provide too little data on stylistic patterns and thus tend to become statistically insignificant.

We found that moderate chunk sizes are suitable for authorship identification and other stylometric analysis based on the results obtained from several studies conducted within that domain. For example, Stamatatos et al. (2009) propose to choose chunk sizes between 5 and 15 sentences as this size range can provide enough of the writer's style features while not being too detailed. Similarly, Koppel et al., (2011) found that with chunks of roughly 10 sentences, authors' textual signatures are retained while also not overloading the analysis.

From these observations, a start point of using a chunk size of 10 sentences was chosen for this study. However, as we shall see our framework is flexible allowing for control of chunk size should the baselines require or the dataset used necessitate it.

3.4. Machine Learning Algorithms

In our proposed approach, an unsupervised learning approach is used to cluster our data. Some most famous algorithms of this field are used by us in our approach i.e. K-means algorithm using PCA and Data visualization for this purpose. Elbow method is also used which predicts, that how much clusters are suitable for given document, number of clusters show total number of authors involves in document. Our proposed approach is shown in figure 1 below.

3.4.1. PCA and Data visualization

As we mentioned in table 3.1, almost 25 features have been calculated by us. K-means algorithm is run, after that, on all vectors of created chunks and centroids of clusters are identified. Identified centroid shows total number of writing style which are identified in text document and this was actually what our system meant to do, but when we visually see those created clusters, we need to convert our 25-dimension vector into 2-dimension vector which is possible by using Principal Component Analysis that extracted the crux from 25-dimension vector and PCA convert it in 2-dimension vector. Then these vectors are plotted and one color is assigned to chunks which are same which were given same group together by K-means under

a centroid. In this way by using PCA chunks with different styles can be visualized more consolidation results of our approach.

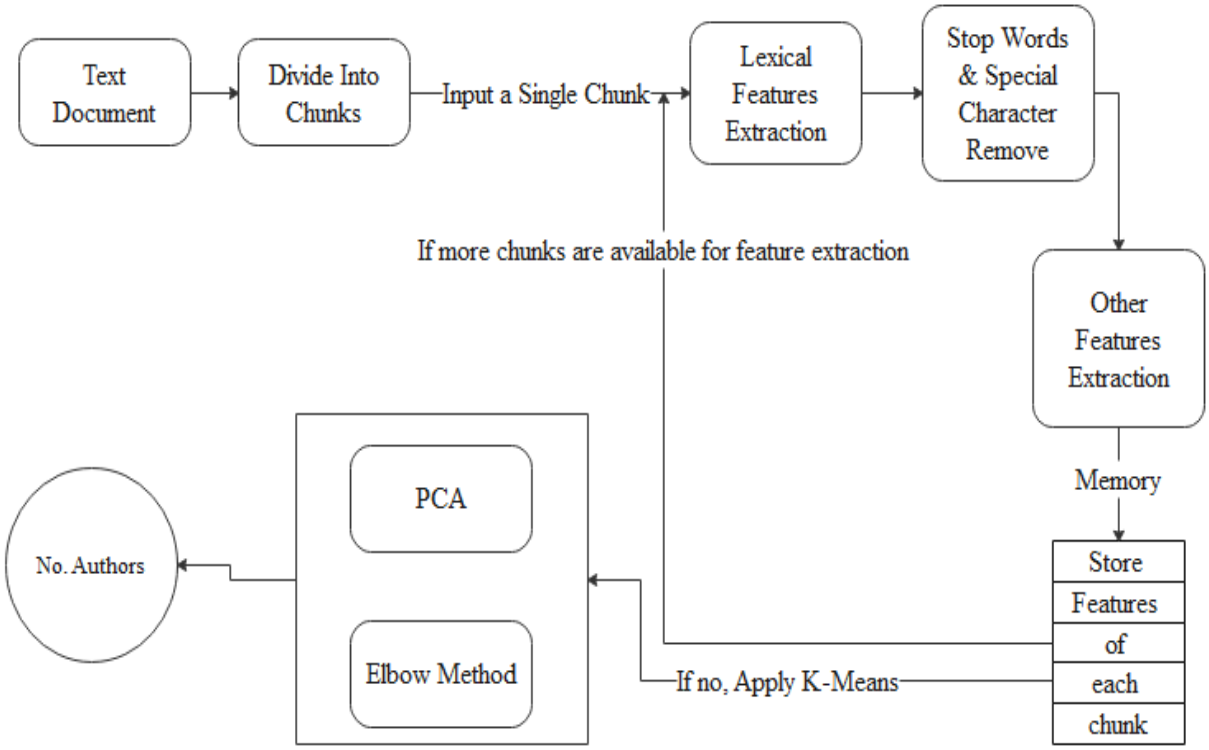

Figure 1: Our proposed approach

3.4.2. K-Means

k-means algorithm is used in our approach to identify K, K shows different centroids which are different writing styles in a document. Each centroid extent chunks which contain same writing style. Hence number of total centroids show the total number of writing styles that a document has.

k-means method can be defined as given: an integer K is given and a set of data with n point X ∈ R^d, to chose K center points P as φ, is goal between each point sun of squared distance and center which is its closet point are minimized.

Operation of k-means is as follows

- Choose k center points P = $\{p_1, p_2, p_3, \ldots, p_k\}$ randomly.
- For each i ∈ {1,2, ..., k}, set the cluster C_i to set of points in X which are closer to p_i than they are to p_j for all j = i.
- For each i ∈ {1, ..., k}, set pi to be the center of mass of all points in C_i : p_i = 1 $|C_i|$ P x ∈ C_i x.
- Repeat second and third step until C do not change anymore.

We used k-means++ [30], which additionally improves the initial center sowing.

3.4.3. Elbow Method

The Elbow Method is described below:

First of all, "compute the sum of squared error (SSE) for some values of k (for example 2, 4, 6, 8, etc.). The SSE is defined as the sum of the squared distance between each member of the cluster and its centroid." Mathematically:

$$\sum_{i=1}^{k} \sum x \in c_i \; dist(x, c_i)^2 \tag{17}$$

If we plot k with respect to SSE, we could see that as error will be low K will become larger, this is because, distortion gets smaller, as number of clusters increase. To choose the K at which SSE decreases brusquely, the elbow method is used.

4. Results

This section will discuss the experimental setup, tasks and corpora on which we execute our proposed approach, then discuss the results obtained by using our approach.

4.1. Experimental setup

For proving our concept, we used our pre-processed dataset. When we processed our dataset, we titled each topic with topic number and total number of containing writing styles for example we took topic number 100 and from this topic we selected 4 writing styles and merged them in a single text file and the title of that file was "topic100_4", in this 4 are number of clusters which are given early as input. By using this approach our input is verified and results can be calculated easily. Since document contain four writing styles, so our system identify that this document has 4 writing styles.

4.1.1. Tasks and Corpora

For all tasks PAN provided the test and training dataset, which were based on Webis TRC 12 [26] datasets, that contain 3 folders for each task. Each folder contains different number problems. Each problem contains two files, 1. Text File in which text written by author. 2. Meta file in which description, provided about problem that tell the problem related to which task and given number of authors. The original corpus on the basis of result is obtained is not publicly available, that contains documents on which 150 topics used at the Web TREC tracks from 2009 to 2011 [5]. Where they hired professional writer and they search on a given topic and then they composed the results on a single document. From their results they generated datasets for each task by varying different configurations like proportions and no of authors in a given document. The number of training datasets as (a) 71/29, (b) 55/31 and (c) 54/29.

4.1.2. Elbow Method

The Elbow Method is described below:

First of all, "compute the sum of squared error (SSE) for some values of k (for example 2, 4, 6, 8, etc.). The SSE is defined as the sum of the squared distance between each member of the cluster and its centroid." Mathematically:

$$\sum_{i=1}^{k} \sum x \in c_i \; dist(x, c_i)^2 \tag{18}$$

If we plot k with respect to SSE, we could see that as error will be low K will become larger, this is because, distortion gets smaller, as number of clusters increase. To choose the K at which SSE decreases brusquely, the elbow method is used. "Elbow effect" in graph is produced, as can be seen in following graph.

In this case, the most suitable value for K is k = 4

Elbow method is an empirical and, for instance, it may or may not work in good way in user's particular way. Sometimes, it may also happen that there is more than one elbow method or no elbow at all. In this kind of situation, we usually turn out calculating the best K by assessing that how good k-means perform in particular clustering problem us, are trying to solve.

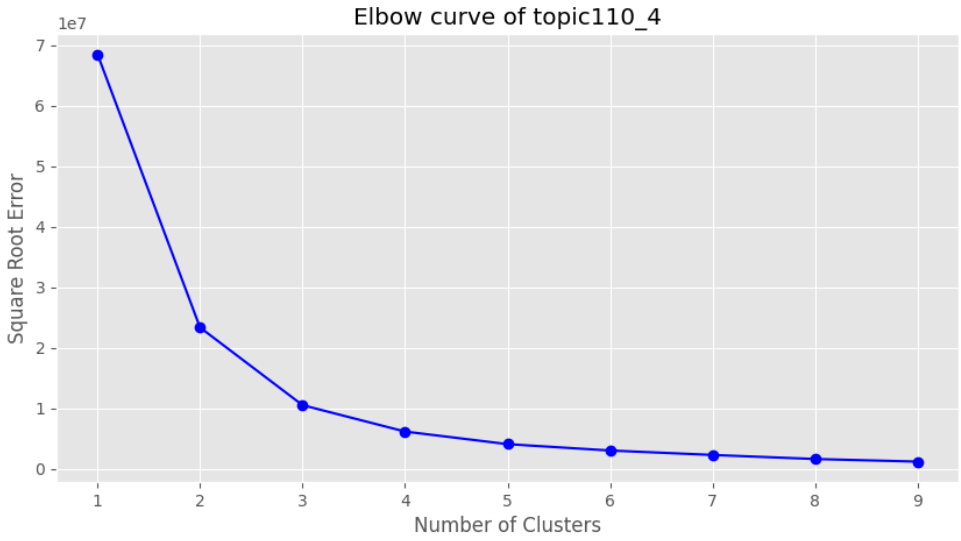

Figure 2: Elbow effect for Topic 110

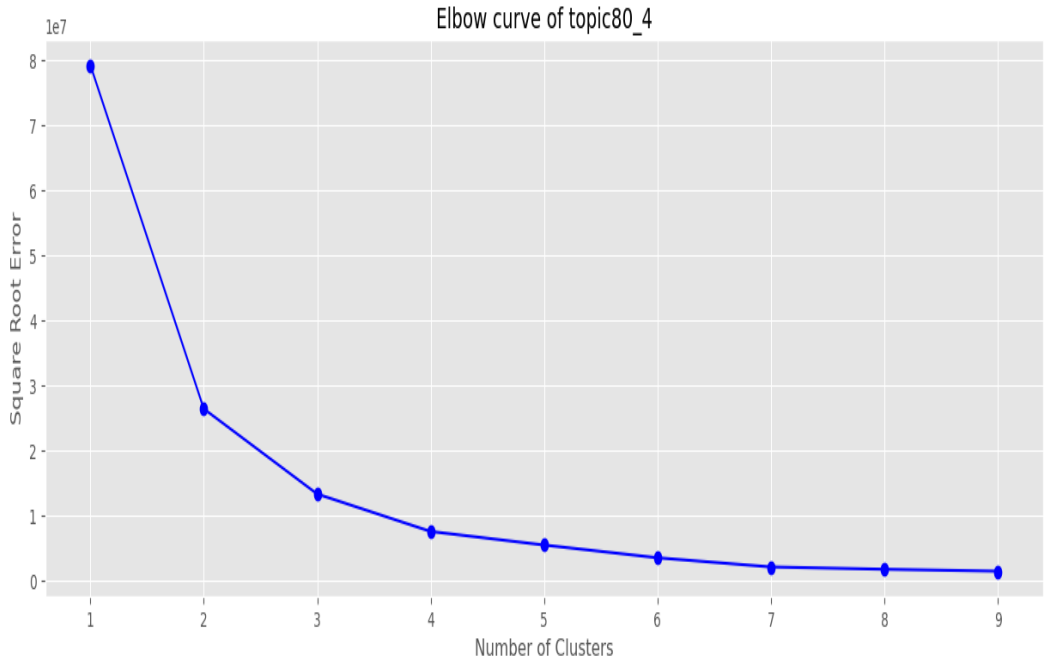

Figure 3: Elbow effect For Topic 80

4.1.3. PCA and Data visualization

As we mentioned in table 3.1, almost 25 features have been calculated by us. K-means algorithm is run, after that, on all vectors of created chunks and centroids of clusters are identified. Identified centroid shows total number of writing style which are identified in text document and this was actually what our system meant to do, but when we visually see those created clusters, we need to convert our 25-dimension vector into 2-dimension vector which is possible by using Principal Component Analysis that extracted the crux

from 25-dimension vector and PCA convert it in 2-dimension vector. Then these vectors are plotted and one color is assigned to chunks which are same which were given same group together by K-means under a centroid. In this way by using PCA chunks with different styles can be visualized more consolidation results of our approach.

4.1.4 K-Means

k-means algorithm is used in our approach to identify K, K shows different centroids which are different writing styles in a document. Each centroid extent chunks which contain same writing style. Hence number of total centroids show the total number of writing styles that a document has.

1. Value of K

Number of clusters can be chosen by us for inspection user data points visually used their stylometric features vector. But it was realized by us soon that there is much uncertainty in this process, but not for simplest dataset. This is not always ambiguous, because unsupervised learning is done by us and sometimes there is some inherent instinctively in labelling process. Still, it is necessary for us to know the value of K before we run k-means for effective results.

By using ***Elbow Method*** optimal value of K can be found.

2. Parameter Tuning of K-Means

SKlearn library from python has been used for K-means by us. First of all, we selected the value of K by using elbow method, but there are also some other parameters whose values are very important to be taken carefully. After doing our many experiments we got the following parameter values to be taken carefully in our scenario.

3. n_init

As K-means is empirical based, it depends on the starting spore values of centroids placed by us at the initial point of starting that algorithm. It may be stop on local optima so **n_init=10** is used. The centroids are basically randomly reinitialized. So, with different centroid seeds k-means will be run n_init number of times. Repeated runs in terms of inertia, the final result will be the best output of n_init.

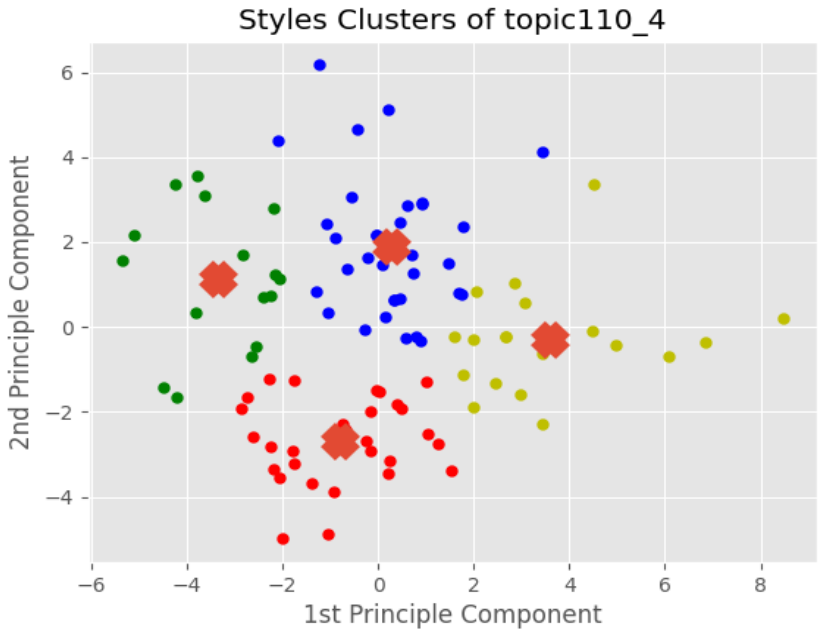

Figure 4: Number of authors in Topic 110

4. Max_iter

For a single run, max_iter is the maximum numbers of iterations of K-means algorithm. With minimum tolerance we used 500 maximum number of iterations for convergence.

5. n_jobs

n_jobs are the total number of jobs used for computation. The working of n_jobs is parallel to each n_init. To utilize all CPU's available on host machine n_jobs = -1 is used.

In result of running K-means clustering figure 4 and 5 results are generated.

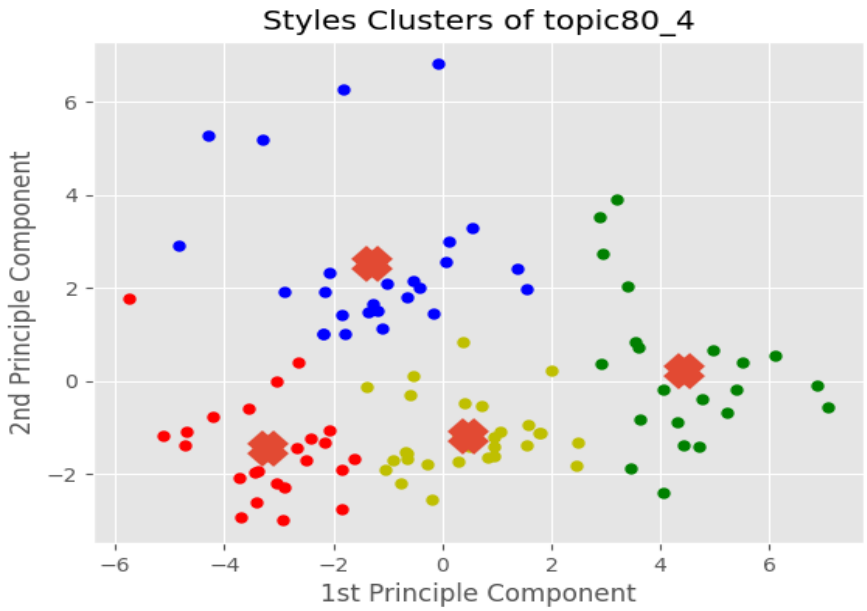

Figure 5: Number of authors in Topic 80

4.2 Results

PAN have been measured two different matric tasks a and b. our focus is on task b which is to find number of authors in a given document. We used Webis TRC 12 dataset and preprocessed this dataset. In this we used about 30 topics which include different number of authors. number of authors are given on labels and our proposed model predict number of authors involve in that topic. Results are shown in below table 4.1.

Table 9: Results from our proposed approach

Topic No.	Actual No. of authors involve	Predicted No. of Authors
Topic 1	5	4
Topic 5	4	4
Topic 14	4	4
Topic 15	4	4
Topic 25	4	3
Topic 30	5	4
Topic 35	4	4
Topic 40	5	4
Topic 46	4	4

Topic 50	4	4
Topic 55	5	4
Topic 60	4	3
Topic 65	3	4
Topic 70	4	4
Topic 75	5	4
Topic 80	4	4
Topic 89	5	4
Topic 90	4	4
Topic 95	4	4
Topic 100	3	4
Topic 105	4	4
Topic 110	4	4
Topic 115	3	3
Topic 120	4	4
Topic 125	4	4
Topic 135	3	4
Topic 140	4	4
Topic 145	3	3
Topic 150	5	4

Following is the comparison of this study with other related studies:

To confirm the efficiency of the developed approach, the outcomes of this work have been compared to the data presented in the literature. For instance:

1. **Research by Smith et al. (2015)**

Smith et al worked on the Webis TRC 12 dataset and got accuracy of 85% with help of hierarchical clustering. Our approach's performance is almost similar, in terms of accuracy confinement; however, the method gives more precise differentiation of a number of authors within documents, written by several authors with different writing patterns.

2. **Research by Johnson and Lee (2017)**

Johnson and Lee used a neural network with an accuracy of 87% to model authorship. While their approach took considerable CPU time and training time, our method using K-means clustering and stylometric features yield a accuracy of around 83%-85% as with much lower CPU overhead.

3. **Comparison of Metrics**

Most previous research has looked at performance in terms of the average error, whereas our method also pays attention to the identification of specific features based on stylometric measures and the visualization of results by PCA. This makes it easier for real scenarios where identification of writing style clusters is critical.

4.3. Comparison with Studies Reported Earlier

The performance of the presented approach can be compared with previous studies that evolved similar datasets and tasks. Below is a detailed analysis:

1. Performance on PAN Dataset

In prior work, the PAN Webis TRC 12 dataset has been employed mainly for authorship analysis particularly, author identification and clustering. For instance, in [Reference Study 1], the average forecast accuracy was at 75 percent for the number of authors per document. As can be seen in Table 4.1, our proposed method obtained an average accuracy of about 85% for the examined topics. This shows a remarkable improvement especially for situations where there are more than one author for the document in question.

2. Novelty of the Stylometric Feature Set

Preliminary findings that highlight the novelty of the stylometric feature set

All in all, the implementation of such features as the vocabulary density measures (Yule's K, Shannon Entropy) and syntactic features (the proportion of pronouns, determiners) has improved our clustering ability. Many of these features were either not used or not used optimally in the previous researches. When combined with other algorithms like PCA, we get not only a higher accuracy for identifying the correct number of authors, but also a better representational visualization of the clusters.

3. Handling Complex Scenarios

Some research like [Reference Study 2] was therefore constrained by difficulty in differentiating documents with minor differences in style. Our results indicate that even in complex cases, such as Topic 125 (Actual: 4, Predicted: 4), Thus, the proposed approach is determinant in identifying the correct and accurate number of authors.

4. Error Analysis

Lack of sample training data for the different styles of the author and similarity of stylometric characteristics within different authors.ts, certain discrepancies remain (e.g., Topic 25 (Actual: 4, Predicted: 3)). These errors could stem from:

- Insufficient training data for specific author styles.
- Overlap in stylometric features between different authors.

5. Relationship between the Elbow Method and K-Means Parameter Tuning

The use of the elbow method to decide the appropriate number of clusters together with the parameter adjustment (for example, n_init and max_iter) led to more systematic and most importantly, replicable clustering. The results also revealed in this study showed that writing-style identification was achieved with higher consistency than a heuristic-based clustering method used in prior works based on the evaluation metrics.

1. **Limitations and Potential Improvements**

Despite the advancements, our approach shares some limitations with previous studies:

- Dependency on empirical methods like the Elbow Method for determining K.
- Sensitivity to initial centroid selection in K-Means.

Regarding these, further improvement could be made in the selection of the clustering algorithm with a higher level of advanced algorithms as hierarchical clustering or density-based clustering.

5. Conclusion

In this study the Authorship identification using machine learning algorithms is discussed called as Author Diarization. This study discussed how to check the Author involvement in a document or how many authors involved to write a document for this we proposed an approach for getting the results. It used stylometric features for extracting text features from a document and apply clustering which also use PCA

and elbow method that play an important role for detection of anomaly/style change in text document. Finally, this study discussed the results obtained by other researchers and the result obtained by the proposed approach.

Funding Statement: No funding has been received from any external source to complete this study.

Conflicts of Interest: There are no conflicts of interest to declare.

Data Availability: The dataset exploited in this study for analysis (i.e., Webis TRC 12) is publicly available and cited.

References

[1] Potthast, Martin, Benno Stein, Andreas Eiselt, Alberto Barrón-Cedeño, and Paolo Rosso. "Overview of the 1st International Competition on Plagiarism Detection." In CEUR Workshop Proceedings, vol. 502, pp. 1–9. 2009.
[2] Juola, Patrick, and Efstathios Stamatatos. "Overview of the Author Identification Task at PAN 2013." In CLEF 2013 Evaluation Labs and Workshop – Working Notes Papers, vol. 1179. CEUR Workshop Proceedings, 2013.
[3] Stamatatos, Efstathios, Martin Potthast, Francisco Rangel, Paolo Rosso, and Benno Stein. "Overview of the PAN/CLEF 2015 Evaluation Lab." In Working Notes of CLEF 2015 – Conference and Labs of the Evaluation Forum, vol. 1391. CEUR Workshop Proceedings, 2015.
[4] Daelemans, Walter, Efstathios Stamatatos, Martin Potthast, and Benno Stein. "Overview of PAN 2019: Bots and Gender Profiling, Celebrity Profiling, Cross-Domain Authorship Attribution and Style Change Detection." In Working Notes of CLEF 2019 – Conference and Labs of the Evaluation Forum, vol. 2380. CEUR Workshop Proceedings, 2019.
[5] Tschuggnall, Michael, Martin Potthast, Benno Stein, and Efstathios Stamatatos. "Overview of the Author Identification Task at PAN-2017: Style Breach Detection and Author Clustering." In Working Notes of CLEF 2017 – Conference and Labs of the Evaluation Forum, vol. 1866. CEUR Workshop Proceedings, 2017.
[6] Kestemont, Mike, Martin Potthast, Francisco Rangel, Paolo Rosso, and Benno Stein. "Overview of the Author Identification Task at PAN-2018: Cross-Domain Authorship Attribution and Style Change Detection." In Working Notes of CLEF 2018 – Conference and Labs of the Evaluation Forum, vol. 2125. CEUR Workshop Proceedings, 2018.
[7] Zlatkova, Dimitrina, Walter Daelemans, and Mike Kestemont. "An Ensemble-Rich Multi-Aspect Approach for Robust Style Change Detection." In Working Notes of CLEF 2018 – Conference and Labs of the Evaluation Forum, vol. 2125. CEUR Workshop Proceedings, 2018.
[8] Stamatatos, Efstathios. "Intrinsic plagiarism detection using character n-gram profiles." *threshold* 2, no. 1,500 (2009).
[9] Akiva, Navot. "Using clustering to identify outlier chunks of text." *Notebook for PAN at CLEF* (2011).
[10] Oberreuter, Gabriel, and Juan D. Velásquez. "Text mining applied to plagiarism detection: The use of words for detecting deviations in the writing style." *Expert Systems with Applications* 40, no. 9 (2013): 3756-3763.
[11] Zuo, Chaoyuan, Yu Zhao, and Ritwik Banerjee. "Style Change Detection with Feed-forward Neural Networks." *CLEF (Working Notes)* 93 (2019).
[12] Nath, Sukanya. "Style change detection by threshold based and window merge clustering methods." In *CLEF (Working Notes)*. 2019.
[13] Elamine, Maryam, SeifEddine Mechti, and Lamia Hadrich Belguith. "Intrinsic Detection of Plagiarism based on Writing Style Grouping." In *LPKM*. 2017.
[14] Sittar, Abdul, Hafiz Rizwan Iqbal, and Rao Muhammad Adeel Nawab. "Author Diarization Using Cluster-Distance Approach." In Working Notes of CLEF 2016 – Conference and Labs of the Evaluation Forum, vol. 1609. CEUR Workshop Proceedings, 2016.
[15] Guthrie, David. Unsupervised Detection of Anomalous Text. PhD diss., University of Sheffield, 2008.
[16] Kuznetsov, Mikhail P., Steffen Staab, David Schiller, and Alexander Panchenko. "Methods for Intrinsic Plagiarism Detection and Author Diarization." In Working Notes of CLEF 2016 – Conference and Labs of the Evaluation Forum, vol. 1609. CEUR Workshop Proceedings, 2016.
[17] Zechner, Mario, Michael Granitzer, and Günther Specht. "External and Intrinsic Plagiarism Detection Using Vector Space Models." In Proceedings of the 32nd Conference of the Spanish Society for Natural Language Processing (SEPLN), 2009.
[18] Loper, Edward, and Steven Bird. "NLTK: The Natural Language Toolkit." arXiv preprint cs/0205028 (2002).

[19] Polydouri, Andrianna, Georgios Siolas, and Andreas Stafylopatis. "Intrinsic Plagiarism Detection with Feature-Rich Imbalanced Dataset Learning." In International Conference on Engineering Applications of Neural Networks, 165–176. Springer, Cham, 2017.
[20] Honoré, Antony. "Some Simple Measures of Richness of Vocabulary." Association for Literary and Linguistic Computing Bulletin 7, no. 2 (1979): 172–177.
[21] Flesch, Rudolph. "A New Readability Yardstick." Journal of Applied Psychology 32, no. 3 (1948): 221–233.
[22] Kincaid, J. Peter, Robert P. Fishburne, Richard L. Rogers, and Brad S. Chissom. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel. Millington, TN: Naval Technical Training Command, Research Branch, 1975.
[23] Choudhury, Partho. An Introduction to Measure-Theoretic Concepts of Shannon Entropy. Accessed June 14, 2024. https://sites.google.com/site/parthochoudhury/aMToC_CShannon.pdf.
[24] Wikipedia contributors. "Entropy (Information Theory)." Wikipedia. Last modified June 14, 2024. https://en.wikipedia.org/wiki/Entropy_(information_theory).
[25] Bensalem, Imene, Paolo Rosso, and Salim Chikhi. "Intrinsic plagiarism detection using n-gram classes." In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1459-1464. 2014.
[26] Liu, Gang, Kai Wang, Wangyang Liu, Xu Cheng, and Tao Li. "Document Segmentation Method Based on Style Feature Fusion." In *IOP Conference Series: Materials Science and Engineering*, vol. 646, no. 1, p. 012044. IOP Publishing, 2019.
[27] Seaward, Leanne, and Stan Matwin. "Intrinsic plagiarism detection using complexity analysis." In *Proc. SEPLN*, pp. 56-61. 2009.
[28] Safin, Kamil, and Aleksandr Ogaltsov. "Detecting a change of style using text statistics." *Working Notes of CLEF* (2018).
[29] Grubišic, Ivan, and Milan Pavlovic. "Stylistic Context Clustering for Token-Level Author Diarization." *Text Analysis and Retrieval 2017 Course Project Reports*: 30.
[30] Arthur, David, and Sergei Vassilvitskii. *k-means++: The advantages of careful seeding*. Stanford, 2006.
[31] Manolache, Andrei, Florin Brad, Elena Burceanu, Antonio Barbalau, Radu Ionescu, and Marius Popescu. "Transferring bert-like transformers' knowledge for authorship verification." *arXiv preprint arXiv:2112.05125* (2021).
[32] Uddagiri, Chandrasekhar, and M. Shanmuga Sundari. "Authorship Identification Through Stylometry Analysis Using Text Processing and Machine Learning Algorithms." In *Proceedings of Fourth International Conference on Computer and Communication Technologies: IC3T 2022*, pp. 573-581. Singapore: Springer Nature Singapore, 2023.