Accepted Manuscript

An integrated approach for intrinsic plagiarism detection

Muna Al-Sallal, Rahat Iqbal, Vasile Palade, Saad Amin, Victor Chang

PII: S0167-739X(17)32601-8
DOI: https://doi.org/10.1016/j.future.2017.11.023
Reference: FUTURE 3815

To appear in: *Future Generation Computer Systems*

Received date : 29 May 2016
Revised date : 29 August 2017
Accepted date : 11 November 2017

Please cite this article as: M. Al-Sallal, R. Iqbal, V. Palade, S. Amin, V. Chang, An integrated approach for intrinsic plagiarism detection, *Future Generation Computer Systems* (2017), https://doi.org/10.1016/j.future.2017.11.023

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

ACCEPTED MANUSCRIPT

An Integrated Approach for Intrinsic Plagiarism Detection

Muna Al-Sallal[1], Rahat Iqbal[1], Vasile Palade[1], Saad Amin[1], Victor Chang[2]
[1]Faculty of Engineering, Environment & Computing
School of Computing, Electronics and Maths
Coventry University
Email: aa6095@coventry.ac.uk; r.iqbal@coventry.ac.uk; vasile.palade@coventry.ac.uk; s.amin@coventry.ac.uk;

[2]International Business School Suzhou,
Xi'an Jiaotong Liverpool University, Suzhou, China
Email: ic.victor.chang@gmail.com

Abstract: Employing effective plagiarism detection methods are seen to be essential in the next generation web. In this paper, we present a novel approach for plagiarism detection without reference collections. The proposed approach relies on using some statistical properties of the most common words, and the Latent Semantic Analysis that is applied to extract the most common words usage patterns. This method aims to generate a model of author's "style" by revealing a set of certain features of authorship. The model generation procedure focuses on just one author, as an attempt to summarise the aspects of an author's style in a definitive and clear-cut manner. The feature set of the intrinsic model were based on the frequency of the most common words, their relative frequencies in the book series, and the deviation of these frequencies across all books for a particular author. The approach has been evaluated using the leave-one-out-cross-validation method on the CEN (Corpus of English Novel) data set. Results have indicated that, by integrating deep latent semantic and stylometric analyses, hidden changes can be identified when a reference collection does not exist. The results have also shown that our Multi-Layer Perceptron based approach statistically outperforms Bayesian Network, Support Vector Machine and Random Forest models, by accurately predicting the author classes with an overall accuracy of 97%.

1. Introduction

Plagiarism detection and authorship analysis approaches have a long history of attempts to improve their performance in detecting text misuse and identifying the author of a suspicious text. However, despite a considerable work in improving such methods, the performance of these methods is still unsatisfactory in some cases. Detecting the imitation of the language used in a particular piece of text is not a challenge for the innovative recent plagiarism detection techniques. The new research trends in the area involve detecting the ideas, methodologies and findings that are reproduced as new work, without proper credit being given to the original author. The majority of the existing methods were built on a notion that all related information is digitalised. A criticism that has been raised against this assumption revolves around the fact that not all sources are digitalised yet [40]. Consequently, a new class of plagiarism detection tools is currently being researched and developed, termed intrinsic plagiarism detection methods. These methods aim to characterise a writer's style using a history of that writer's existing work, and they do not use a collection of references to compare with [37]. Such methods rely on capturing the variations in the written text by extracting different types of features. Then, a comparison between the suspicious text and the same author's work is performed in order to identify the variation patterns.

The methods of tackling plagiarism were stimulated by the authorship analysis approaches, which use several text analysis techniques to infer the authorship of suspicious texts. In traditional authorship analysis, a suspicious text is attributed to one author, when given a group of authors with their textual samples [31]. The authorship analysis approaches have stemmed from a linguistic root called stylometry, which refers to the field of study that works on quantifying the author's writing style features based on statistics computations [1]. Stylometry relies on a fact that each author has irreplaceable writing habits that cannot be imitated [9].

The procedure of quantifying the most common words (MCWs) in a document is assumed to be the most effective method in stylometric studies and, recently, in intrinsic plagiarism detection approaches [32]. Most linguistics experts have argued that each author has a specific group of MCWs that feature their writing style and are assumed to be closed-class [31]. The procedure of identifying the members of a closed-class MCWs for each author is based on analysing these word usage patterns. Two main factors that support the importance of MCWs as discriminative attributes are their sub-conscious usage and also their independency from the corpus

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

topics. It is argued that there is no one author who can write different documents using different usage patterns of the MCWs [7].

This paper presents a new method for intrinsic plagiarism detection, where no reference available to compare with is available. The method relies on the integration of four well-known techniques: bag of words (BOW), latent semantic analysis (LSA), stylometry and multilayer perceptron neural networks (MLP). These techniques were used to enhance the quantifying procedure of the implicit stylistic features of the text. To the best of our knowledge, the proposed method is the first one which applies this combination of techniques to capture the text variation between two documents. This method targets the authorial attributes and ignores any content related topic. The core component of this method is stylometry, which relies on deriving sets of features based on MCW frequencies. The performance of this method will be measured based on how the derived sets of features perform using MLPs and other machine learning algorithms. BOW and LSA are used in the pre-processing stage. BOW has been used as a first step for feature generation based on MCWs, while LSA was used as a mean of shrinking the vector's dimensions. The experimental design relies on using the Corpus of English Novels (CEN) dataset. The CEN dataset contains 292 novels that were written by 25 British (including Irish) and North American novelists. The novels were written in the period between 1881 and 1922, furthermore all authors were born between 1848 and 1963 and represent roughly one generation of writers. The dataset was composed by Hendrik De Smet. The data has been used in many studies as an example of sample texts from different authors. The leave-one-out-cross-validation (LOOCV) method was used as a sampling technique for training just the target author books. The books of the target author are trained as positive examples based on one-class classification roles in order to generate the prediction model.

The work presented in this paper is part of a wider endeavour towards developing intelligent plagiarism detection tools as well as providing awareness on plagiarism to the academic and research communities. The rest of the paper is organised as follows: Section 2 presents a literature review. Section 3 presents a general discussion on the selected models and methods which are used as part of the proposed approach. Section 4 presents the proposed integrated approach and discusses the various phases of the approach as well as its implementation details. Results are presented in Section 5. Section 6 concludes the paper by summarising the main contributions of the paper and outlining the future work.

2. Previous work

The current literature has split the plagiarism detection methods into two forms: extrinsic and intrinsic [5]. Extrinsic plagiarism detection methods rely on comparing the suspicious document or string of text to a body of known, classified documents [5]. While these methods perform well to some extent for copy and paste misconduct, the detection assumption was built on a notion that all related information was digitised. It is argued that not all sources are digitised, hence, a new class of plagiarism-detection tools named as intrinsic detection methods [39]. A comparative study to evaluate the state of the art in plagiarism detection is reported in [17]. In their study, they highlighted that the most significant challenge in plagiarism detection field was to identify the text author. They also recommended that incorporating stylistic variation detection techniques with the current plagiarism detection approaches can substantially enhance the plagiarism detection performance.

For centuries, scholars have sought to find more reliable ways to prove the authorship of certain important documents. Even scholars who have spent a lifetime analysing certain documents and authors often did not agree on authorship (e.g., a number of works generally attributed to Shakespeare are argued to have been written by Marlow instead [37]. Intrinsic plagiarism detection method can help to generate a model of the author's style and help reveal certain features of authorship (e.g., for literary analysis). However, these methods are normally evaluated based on a small dataset [23]. The early basic set of techniques in authorship analysis has relied on selecting features from an author's written texts that are unique to that author (unicity) and these features do not change over time (invariant). These techniques were discussed and defined in the late 19th century by Mendenhall (1887), who studied the texts of Shakespeare, as well as Marlow and other contemporaries [24]. Mendenhall ultimately discovered that a characteristic can often be found by plotting the curve of frequency vs. word length for a particular author. These two characteristics have established a foundation for the characterisation of an author's writing style and formed a strong basis for statistical approaches [32]. Researchers have continued to search for a single feature that is unique for a specific author and unchangeable during the time. Many suggestions for such features were established, such as calculating the average word length suggested by [15]. Another suggestion by [35], to calculate the average number of words in each sentence, was proposed. These calculations and their counterparts are considered insufficient to identify text authors [19].

ACCEPTED MANUSCRIPT

The next most sophisticated development in the history of stylistic methods represents the next generation class of features originated with the most common words (MCW), which can sometimes be called function words [31]. Function words are language elements without (much) inherent meaning, whose primary purpose was to clarify the relationship between words' classes in different textual parts. An analysis of these parts and the statistics of the most common words have remained a popular topic ever since its inception [25]. Mosteller and Wallace (19640) investigated the authorship of 146 political articles written by James Madison, Alexander Hamilton and John Jay [25]. This was an important milestone work in the field of authorship analysis. This issue was named The Federalist Papers disputation, as twelve of these articles were claimed to be written by Madison and Hamilton. The study found that measuring the frequencies of a specific set of the most common words could result in improving the prediction of the text author, compared to content words or other word classes. One logical explanation for the outcomes of this study is that the unconscious use of a set of words remains constant, even when the topic changes. The study of Mosteller and Wallace was assumed as a solid foundation for using statistics in authorship analysis. In addition, the application of the study has led to the birth of stylometry. This paper proposes approaches which were originally inspired by statistical analysis approaches that were conducted by [25].

Holmes (1998) confirmed that the use of the most common words is more effective in distinguishing between authors as each author has a unique usage pattern of this class of words [34]. This hypothesis was also confirmed by several authors [16] [20] [26] [38]. An important application on using the most common words was adopted by [8], who applied principle component analysis (PCA) on a set of the most common word frequencies. PCA was able to connect a wide range of measures and project them onto a space to measure the similarity distance between several authors [8]. This trend in research encouraged other researchers to follow Burrow's procedure. Biber (1995) applied a statistical method to describe variability among features [6]. The method was called factor analysis. The author has used this method to discriminate between four texts' languages.

The advent of machine learning algorithms in the authorship analysis research field influenced the research movement. Multilayer perceptron, a basic artificial neural network algorithm, was employed by [34] for the authorship analysis task. Tweedie and his colleagues used three hidden layers to train the political articles with a conjugate gradient and two output layers [34]. They reported that the results were harmonised with the previous studies on the same articles. Support Vector Machines (SVM) was introduced to recognise the stylistic features of seven authors by [13]. The dataset includes 2,652 newspaper trainings written by several authors covering three subjects, with a detection accuracy ranging from 60% to 80%.

Depending on the previous studies, a textual features taxonomy has been developed for authorship analysis tasks by [38]. Figure 1 presents four types of feature sets; each set includes a group of influential features that can affect the performance of detection approaches. The textual features types are: lexical, syntactic and structural and content specific. Structural features include paragraph length, use of signature and specified indentation. These features were considered as discriminative authorial attributions for the authors' writing style. Such features strongly depend on the person's writing habits. Lexical features include the frequencies of any class of words based on the predefined task and also the punctuation frequencies, as shown in Figure 1.

The syntactic features can be defined by the function words' usage, punctuation usage and part of speech (POS) usage. Finally, the content specific features are the words that are related to a specific domain and the keyword frequencies. Figure 2 presents the feature types and gives examples for each type.

ACCEPTED MANUSCRIPT

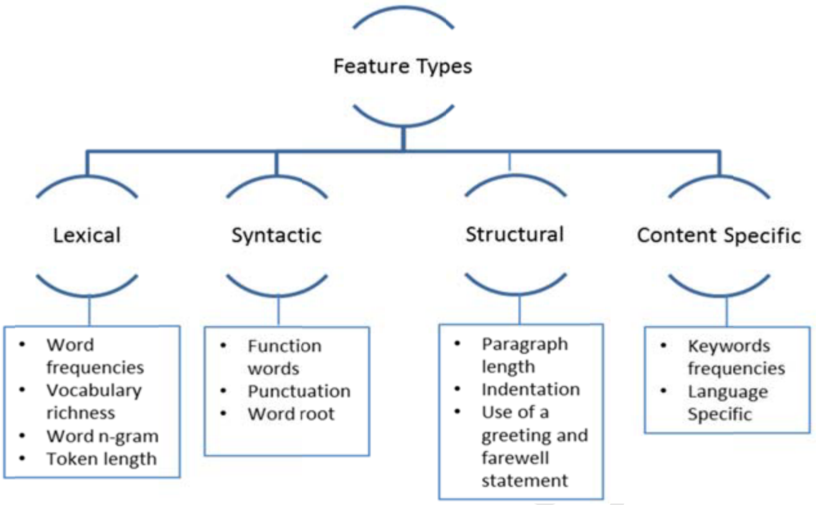


Figure 1: The features types' taxonomy and the most important related features; this figure was inspired by the study of [38]

Koppel and Schler (2004) proposed a one-class classification method in order to identify if a specific written text was written by a target author or not [19]. The method works by stemming two pieces of text and then analysing them using computational stylistics. Based on the analysis, it decides if these two texts were written by one author or more. They concluded that the use of negative examples in the language model influenced the classification accuracy. In the same context, they further identified three types of scenario for their approach to be performed. They proposed a classification procedure to detect the author of a text when there is no candidate corpus, so they analysed the writing based on age, education level, and gender and so on. In the second scenario, they assumed there is a large number of authors (thousands) and the available sample of text for each is very scarce. Koppel's group described this scenario as searching for a "needle-in-a-haystack". In the third one, which the assumption is based on, there are no closed references set to compare with but there is one suspicious set. This is called authorship verification or intrinsic plagiarism and the challenge is to decide if the suspicious text is the author's or is not. They concluded that Bayes and the Support Vector Machines (SVM) performed better in the context of their experiment.

[39] applied the stylometric method based on the average number of sentence lengths for all documents and the number of word classes such as nouns, adjectives. They also calculated the frequencies of special words (frequent and rare words) and their average frequencies. All features were extracted from both the suspicious and original documents to form the input variable sets for machine learning. It is also stated in [39] that the feature sets were analysed using SVM for classification tasks. They reported that the average word frequencies and the average sentence length outperformed other feature sets.

A subsequent study by [40] used the same feature sets from their previous study to investigate the performance of their approach by measuring the vocabulary richness. They calculated the vocabulary richness by dividing the average word length by the sentence length. They used a dataset of fifty documents written in German that were artificially partially plagiarised and then employed them in a linear classification algorithm. They followed the method that was applied by [38] by applying a feature combination instead of an individual set.

A major drawback of intrinsic plagiarism detection methods and their parent class of stylometric methods is that they rely on such a small training data set. Typically, the number of documents available for an author under suspicion of plagiarism is fairly small. Style is dependent upon not just the author, but also the level of technicality expected from the work, the length of the work, purpose, and degree of formality. Use of the first person subject and imperative tense would affect a persuasive essay; however, this would not be the case for a scientific report. The small sample size plagues all stylometric methods, and, likewise, all intrinsic plagiarism detection methods.

In order to overcome the small sample size data set problem inherent to the stylometric method style of analyses, researchers have proposed various kinds of text and document features, as shown in Figure 1. These range from simple (tokens such as word length, word per sentence, and other distributions), to higher level

ACCEPTED MANUSCRIPT

(syntactic features such as frequency of the passive voice, nominalization count, and distributions of frequency of different parts of speech tags), and to expert-based (measures of vocabulary and rare word richness). [39] proposed the use of more standard and word-frequency features to develop a plagiarism-detection method called a "taxonomic tree". The tree presents the taxonomy of the misconduct, as shown in Figure 2. This also shows the specific part of plagiarism detection without the available corpus to compare with.

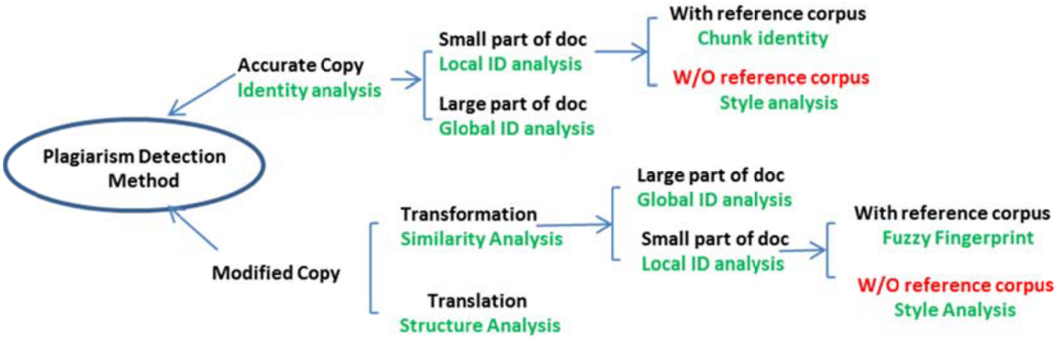


Figure 2: Taxonomic tree of plagiarism-detection methods according to reference document collection size, style of text analysis, and stage in the plagiarism detection process (i.e., processing of accurate copy vs. modified copy) [30].

Stylometry was defined as the linguistic root for assigning a text to its reliable author. This method is based on a statistical analysis of an author's writing style [12]. The textual feature types include lexical, syntactic and structural, and content specific; however, each study can have different feature types based on its aim and experimental corpus.

3. Background

3.1 Bag of Words (BOW)

BOW is one of the popular text representation techniques that is used to represent text in many applications, in particular text classification. BOW relies on a notion that each word establishes a dimension in a vector space isolated from any other words [29]. Researchers have reported that the BOW method can perform much better when integrated with dimensionality reduction methods. This recommendation was proposed because BOW follows a strategy that each word has its own representation in vectors space with no connection to other words [4]. The functionality of BOWs is to break documents into unique words, counting the frequency of each term to form the "baseline" features'. Word counts are important because they form the basic input for a common class of text classification technique.

3.2 Latent Semantic Analysis (LSA)

Latent Semantic Analysis (LSA) is a technique that captures latent semantic associations based on the usage of words [22]. This method has been used as an information retrieval technique (Edmunds 1997), and lately for plagiarism detection (Cosma, 2008; Ceska, 2009). It works in deriving measures of the similarity of meaning between words from the text to mimic human word sorting and category judgments. LSA does not use any linguistic elements, such as grammar, syntactic parser or dictionaries. However, it depends on parsing the raw text into words and it works on extracting the semantic associations based on a pure mathematical model called Singular Value Decomposition (SVD) [22].

3.3 Stylometry

Stylometry relies on the statistical use of computational algorithms to analyse the writing style of a specific author. This method is used to uncover the variation between two pieces of texts and assumed each author has an inimitable writing practice that is conducted unconsciously. These inimitable writing practices can be computed to create a unique writing style for each author in order to compare with others. These unique writing features are measured to create an author profile against which other texts or authors can be compared [3]. It is a well-known method and is widely used in different applications, such as forensic analysis and authorship studies to assign a piece of text with evidences to a specific author based on stylometric quantification.

A study conducted by [27] has claimed that stylometric analysis is considered to be one of the most trusted procedures in recent years. They also argued that stylometry can be used to analyse the authors writing styles and extract informative features for best authorship analysis practices.

A unique word usage pattern can be captured for each author to generate a signature recursive pattern for the text. One of the most stylometric salient features is to use the frequency of function words to quantify stylistic features [25] [9] [33]. The use of function words is considered to be the best discriminant approach in stylometric methods owing to its independence from topics and its unconscious use by authors [7] [33]. Currently most researchers depend on computational processes instead of linguistics procedures [33].

3.4 Machine Learning Techniques for Classification and Feature Selection

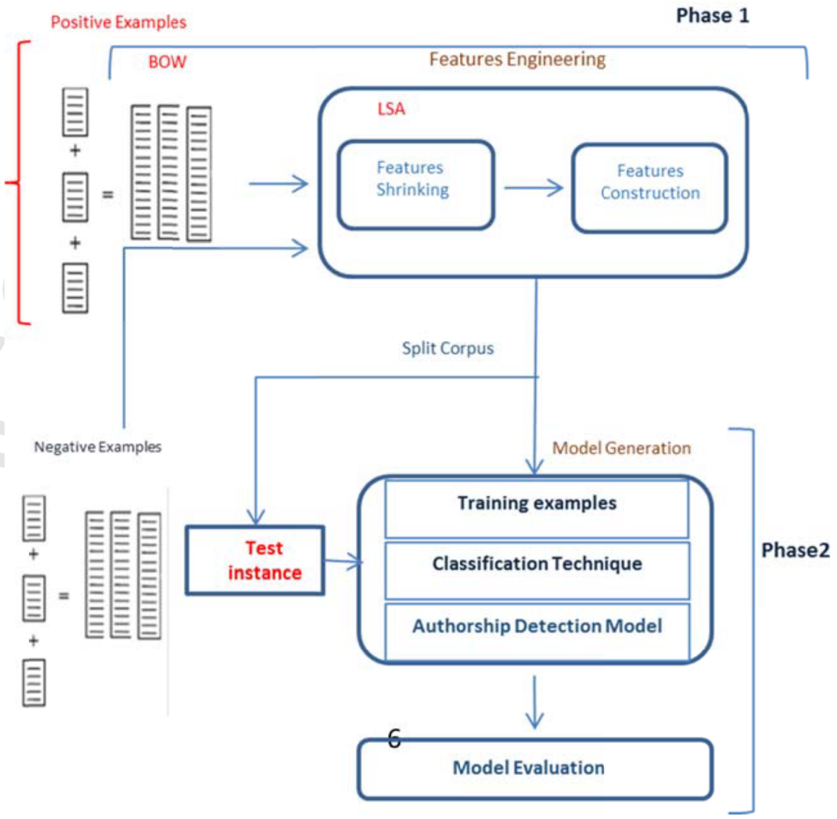
A great variety of machine learning algorithms and techniques can be used, and a detailed review of them here is beyond the scope of the present work. However, probably the most important role of machine-learning methods for our work is in the selection of features. A machine learning model is "only as good as the data put in" [26]. Machine learning methods can pick well from among wide ranges of feature types, in order to generate features for training the models. In addition, they can help generate features of their own, such as character n-grams, function words (of, the, to, other prepositions, etc.), etc. The choice of machine learning method to be used is also of great importance to the ultimate success of a model developed for a certain problem.

4. Proposed Approach

This section discusses the proposed approach for intrinsic plagiarism detection, as shown in Figure 3. The approach is based on four well-known techniques, namely, Bag of Words (BOW), Latent Semantic Analysis (LSA), Stylometry and Multi-Layer Perceptron (MLP). This approach has two main phases and several steps, as shown in Figure 3 and described below:

Phase 1: This phase deals with text representation and features preparation for the second phase, and it has three main steps:

a. Creating bag of words by using the most common words as content-free features.
b. Applying LSA to shrink the high dimensional vectors space that resulted from BOW. LSA application is limited to work as a dimensionality reduction mean.
c. Deriving the proposed features sets from the MCW frequencies.



ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Figure 3: Proposed model for intrinsic validation using Stylometry and LSA

Phase 2: This phase includes the following steps:

a. Applying LOOCV method as a data sampling method for training just the target author books as positive examples in one-class technique.
b. Applying MLP as a classification algorithm to train the positive examples of the target label books based on the derived feature sets.
c. Then a text example is used to generate the prediction model.

This method has relied on the derivatives of the MCW frequencies to build the prediction of the authorship model. The four proposed sets of statistical features are derived in order to capture each author's writing styles patterns:

- The frequencies of the MCWs;
- The relative frequencies of each MCWs;
- The in-series proportional frequencies (e.g., 2nd / 1st, etc.); and
- The z-scores, i.e., a statistical measurement that count the number of standard deviations above and below the mean.

The third and fourth sets of features mainly rely on estimating the probability from adjacent words. This kind of estimation has played a strong role in disclosing important connections between MCWs and exposing their usage patterns. A number of different machine learning models were generated and trained using the features described above. It was desired that a representative of each major class of machine learning methods be used. Therefore, a neural network (multilayer perceptron - MLP), a Bayesian network (BN), a support vector machine (SVM), and a random forest (RF) were all generated, one for each of the 4 word frequency schemes.

With regard to the application of the proposed techniques, one-class classification was employed to conclude if the test document was written by the target author (the author that was trained). The specific author's documents form the positive training set, while the negative examples are anonymous without specific labels. Leave one-out-cross-validation (LOOCV) is used in order to switch the roles between testing and training data for full insights in the prediction performance evaluation.

4.1 Implementation of the Proposed Approach

The following subsections discuss the main components and their implementations in the proposed intrinsic method of plagiarism detection.

4.1.1 Bag of Words (BOW)

The BOW method breaks a document into all of its unique words and count the frequency of each word. In the field of stylometry, the most common words (MCW) are salient elements. This approach is entirely reliant on the most common words frequencies. We did not use content words or any other linguistic elements, except a set of MCW frequencies representing all the books. The bag of words which created for this approach includes MCW and their frequencies.

Implementation of BOW

As an initial step, the BOW model breaks documents into MCWS words, counting the frequency of each word, forming the baseline for each author's documents. Each document in the CEN dataset, which contained sets of books for 25 authors was represented by a BOW, so each author had several BOWs based on the number of documents in the author's dataset. The following points clarify the first phase procedure.

1. The MCW frequencies are calculated, and then each book has its own list.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

2. After the frequencies calculation for all books, the (N x M) matrix, where N is the number of books and M is the number of MCWs, is constructed. The frequency matrix can be expressed as follows.

$$Fr = \begin{pmatrix} fr_1 & \cdots & fr_{1N} \\ \vdots & \ddots & \vdots \\ fr_{M1} & \cdots & fr_{MN} \end{pmatrix} \qquad (1)$$

Each column represents the most common words in each book, and each row represents the distribution of a specific word in all books, which can be denoted by $\boldsymbol{f_{w/b}}$. The BOW method has been used to generate an initial feature set using just content-free words. This method works on capturing the usage of writing stylistic features to identify the text authorship without the use of content features. This represents phase (1) as pointed out in figure 3.

4.1.2 Features Engineering (FE)

This component includes two sub-components; features shrinking and features constructing. It is known that most common words have high frequencies in the text; however authors have their special set of MCW. To reduce the vectors space of MCW that resulted from BOW, LSA has been used for feature space shrinking. The goal of using LSA is to capture the usage patterns of MCW for each author. Another sub-component is features construction. Researchers have demonstrated the ability of LSA in capturing the transitivity relationship between features' words. An example of transitivity (order co-occurrence) is shown in Figure 4, which simplifies the transitivity correlations using Doc2, Doc2 and Doc3 as an example. Doc1 contains word A and word B and connects with Doc2 by word B that occurs in both documents. Doc2 in turn connects with Doc3 by word C, as a result word A connects to word C by word B, as a sequence word A connects to word E as a 3rd co-occurrence level. After text transformation using BOW and LSA, another pre-classification step was applied and feature sets based on MCW were defined.

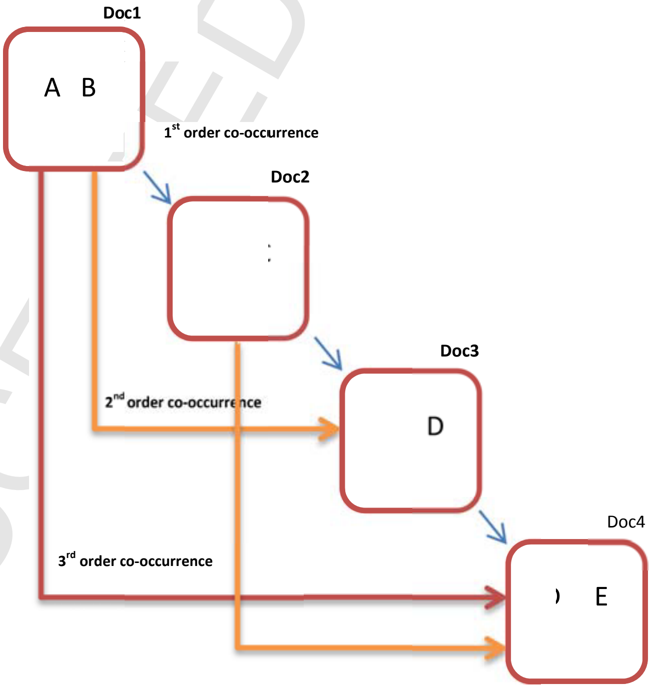


Figure 4: An example of the order co-occurrence tracing

The process of feature selection is assumed to be a key issue in most applications including authorship analysis. The proposed derivative features for this method have been selected based on the baseline frequencies. Once these baseline frequencies were calculated, their frequencies were expressed as a proportion of the total word count in the book. Then, the words were re-expressed as proportions in-series, in-series proportional frequency being represented by (2nd / 1st, etc.). Finally, the z-value of the variance respective to the overall set of documents for the author, are all calculated, for each book. The z-score measure was used as a mean for predicting significant changes in MCWs usage. Equation 4 shows the calculation procedure of z-score by using the mean and standard deviation formulas. The importance of z-score is to show the abnormality behaviour of most common words between the test book and other books of the target author.

To clarify, if zw/b < -1 the MCW appear less frequently in the tested book than its usually distribution in other books in this class than other classes. Furthermore, if zw/b > 1 means the MCW appears more frequently in this class over other classes. The idea is to measure the distribution of each class of books. The significance of the above metric is supported by the notion that analysing internal text structure can enhance the process of capturing variance patterns. Due to the specificity of the intrinsic method for detecting plagiarism sets of features that rely on the raw frequencies of most common words were proposed.

Implementation of FE

As explained above, an innovative feature engineering method was developed and applied in two steps.

Step 1: Co-occurrence feature extraction using LSA

Latent Semantic Analysis (LSA) was employed to extract the co-occurrence feature matrix from each author dataset (25 datasets) and reveal the author's specific patterns based on MCWs. It also directly models the relationship between MCW on the basis of the usage they share. The application of LSA results in the construction of the statistics features matrix.

Step 2: Feature construction

The second step is features construction, where statistical groups of features have been devised including: frequencies (as a proportion of total words), in-series proportional frequencies (2nd / 1st, etc.), and the z-value. The z-score $\boldsymbol{(z_{w/b}}$ is calculated using the mean $\boldsymbol{\mu_w}$ and standard deviation $\boldsymbol{\sigma_w)}$**.**

$$\mu_w = \frac{1}{N}\sum_{b=1}^{N} fr_{w/b} \qquad (2)$$

$$\sigma_w = \sqrt{\frac{1}{N-1}\sum_{b=1}^{N}(f\, r_{w/b} - \mu_w\,)^2} \qquad (3)$$

$$z_{w/b} = \frac{fr_{w/b} - \mu_w}{\sigma_w} \qquad (4)$$

Such calculations express the deviation of the MCW frequency in each book when compared to the corpus average. In other words, the z-score means the process of measuring the abnormality behaviour of MCW frequency with regard to the corpus statistics.

4.1.3 Authorship Generation Model (AGM)

The traditional classification process includes documents with labels, each label belonging to a specific class. For this research, a method was proposed to facilitate the classification process for an individual class of documents [18][28]. A set of documents was labelled to a specific class, which was named as positive examples. In order to enhance the classification process, outlying samples were generated which represent the negative examples. The positive samples represent the class of the target author and all the samples from this class are trained. For the negative examples, training is performed for all other author samples without identifying the

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

labels of the classes, so all other author's books are labelled to be tampered. Figure 5 describes the mechanism for the proposed classification procedure that was used in the intrinsic method for plagiarism detection. The training method of this method differs from the traditional classification method as just one-class examples are trained.

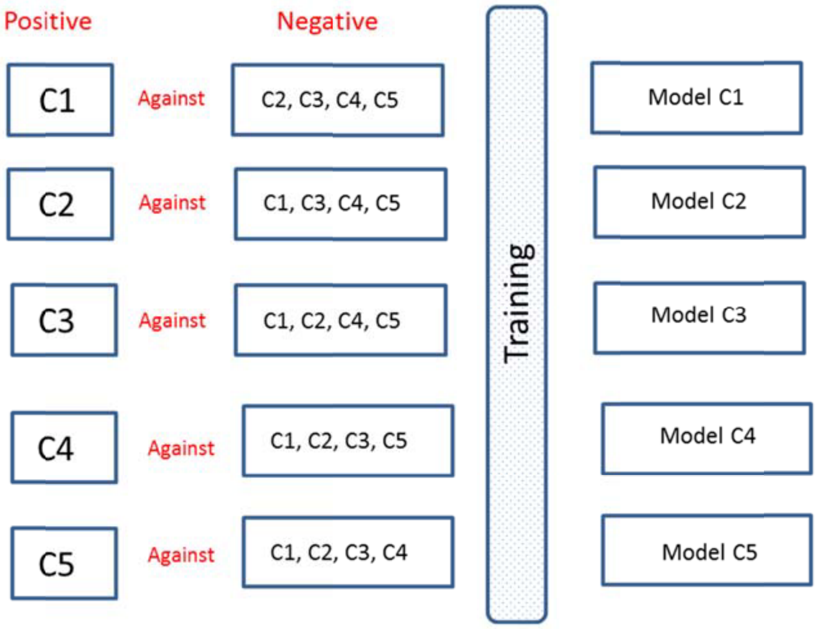


Figure 5: The classification method that was adopted in the proposed intrinsic method (Tax, 2001).

Implementation of Authorship Generation Model

The one-class classification procedure is implemented in order to train separately the data description for each class, which is called the target class. The feature sets of the target class are considered as positive examples and all the data for the other classes are considered negatives (outlier data). The training procedure was performed on the target author documents that were already labelled by their names, so there is just one labelled documents to be trained. In order to apply an efficient classification process there is a need for another type of information just to balance the classification. This information is called negative examples and they do not need to have a specific label. The classifier needs to decide if this text belongs to the target author (tampered-free) or not (tampered).

It is worthy of note that the target author's books represent the unique source of quantifying the author's stylistic features which will be used to build a model that can be used to track variation in the writing style. The other books which are named the negative examples were used to generate the abnormality against the target class as depicted in figure 4. A multilayer perceptron (MLP) learning model was generated and trained using the features described above.

MLP was applied to predict whether or not the anonymous book belongs to the target author based on the one-class classification procedure. The MLP algorithm is a feed-forward ANN (artificial-neural-network) model which includes several layers of nodes. The back-propagation technique is used for training the network. It was used the 'backpropagation' algorithm from Weka 3.7. The MLP parameters were set based on initial empirical attempts; the number of epochs was set to 500, the rate of learning was 0.3, the momentum value was 0.2 (i.e., the default value).

ACCEPTED MANUSCRIPT

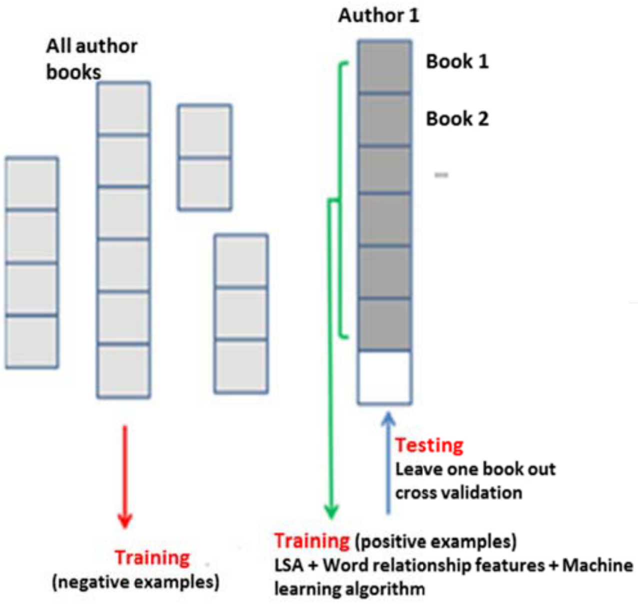

Author 1
All author books
Book 1
Book 2
Testing
Leave one book out cross validation
Training (positive examples)
LSA + Word relationship features + Machine learning algorithm
Training
(negative examples)

Fig 6: Positive examples, for each training set, consist of books for the particular author, while negative examples consist of all works not belonging to the author.

Figure 6 represents the method of using positive and negative examples; the positives consist of books for the particular author. While negative examples consist of all works not belonging to the target author. The model of authorship (the intrinsic method) for detecting plagiarism was generated based on BOW, LSA, Stylometry and the MLP algorithm which were integrated with innovative features composition. An iterative learning and testing procedure was applied using the leave-one-out-cross-validation (LOOCV) technique to develop a robust authorship detection model. The procedure of (LOOCV) was applied for each book in the target author dataset to train the model. Figure 7 describes the mechanism that was applied in order to perform LOOCV sampling technique, with the blue squared shapes representing the training examples, while the red ones represent the test examples that are used to validate the model.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

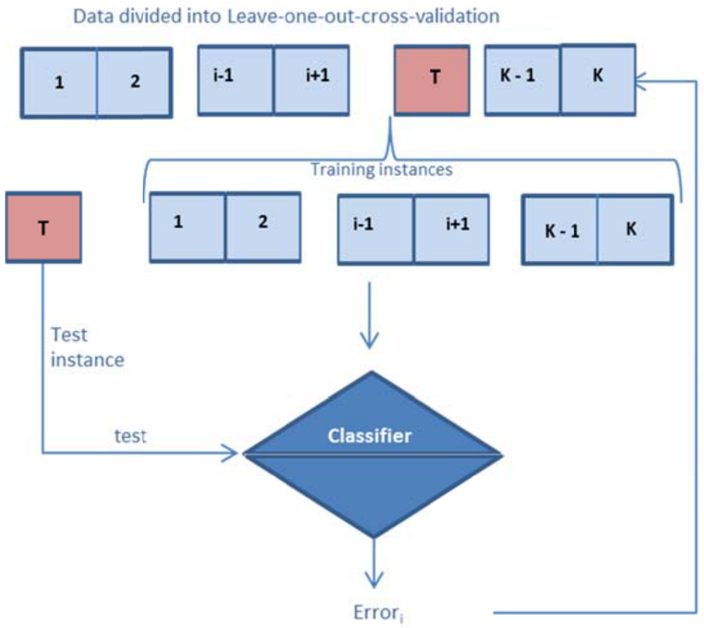


Figure 7: Cross section for leave-book-out-cross-validation method

5. Evaluation of the Results

Several evaluation measures were used, namely, sensitivity, specificity and likelihood ratio. The likelihood ratio is an independent metric which works by putting more confidence on the results and weakens the error potential. This calculation can be performed by applying the formula in Equation 5 to rule-in that the text is plagiarised. While the formula in Equation 6 can be used to rule-out that the text is plagiarised. Another metric, named the confidence interval metric, was used to express the reliability and validity associated with a proposed sampling method. The confidence of the classifier prediction performance can be defined as an indicator of the reliability of the detection results [2]. In other words, the confidence interval represents how precise and stable are the performance measurements when the experiments are repeated again. The confidence interval metric was used at the individual books level in order to assess the performance of the proposed approach on each author set. On the other hand, the negative LR^- (likelihood ratio) as shown in Equation 6 was used to assess the approach performance on all authors' datasets. Both metrics enhance the credibility of the method on an individual book basis as well as entire author's dataset.

$$\text{Positive Likelihood Ratio } (LR)^+ = \text{Sensitivity} / (1\text{-Specificity}) \quad (5)$$
$$\text{Negative Likelihood Ratio } (LR)^- = (1\text{- Sensitivity}) / \text{Specificity} \quad (6)$$

Table 1. The standard confusion matrix

	Positive Class	Negative Classes
Classified as Positive	True Positives **TP**	False Positives **FP**
Classified as Negative	False Negatives **FN**	True Negatives **TN**

ACCEPTED MANUSCRIPT

The confusion matrix in Table 1 describes the process of classification on a set of test data for which the classes are identified. The primary parameter adjusted across a range of values in order to explore predictive capability was the frequencies of MCWs.

Tables from 2 to 5 present the prediction performance of four corpora (authors) as sample indicators of the prediction performance of the proposed approach.

Table 2: The prediction results on the "Gertrude Atherton" dataset

Book	Prediction correct	Confidence (0-1)
1906 rezanov	+	**0.968**
1900 senator north	+	0.936
1921 the sisters-in-law	+	0.948
1922 sleeping fires	+	**0.98**
1888 what dreams come	+	**0.95**
1902 the splindid idle forties	+	**0.969**
1918 the white morning	+	**0.978**
1898 the valiant runaways	+	**0.98**
1900 the doomswoman	+	0.945
1919 the avalanche	-	0.884

In Tables 2 to 5, a "+" sign in the "Prediction correct" column indicates a correct prediction of the author, while the confidence of the prediction made (whether correct or not; for incorrect predictions this is the confidence held that the incorrect prediction was in fact correct) is indicated in the rightmost column. The confidence interval shows the level of credibility on the prediction results and is used to infer that the true value lies between the determined two points. Most studies rely on the 95% confidence interval interpreted to be occurred between the values (0-1) (Field, 2013).

The tables with the prediction results presented books that scored higher, smaller or equal to 0.95 confidence level. If repeated samples were taken and the 95% confidence interval was computed for each sample, then the performance of the proposed approach can be described as 95% generalised to real-world samples. The confidence-level values express that the prediction ability of the intrinsic plagiarism proposed approach is reliable and the proposed approach is likely to get good performance on other samples.

Table 3: The prediction results on the Henry Seton corpus

Book	Prediction correct	Confidence (0-1)
1897 in kedar's tents	+	**0.955**
1900 the isle of unrest	+	0.942
1892 the slave of the lamp	-	0.887
1894 with edged tools	+	**0.953**
1902 the vultures	+	0.948
1892 from one generation to another	+	**0.968**
1901 the velvet glove	+	**0.956**
1895 the sowers	+	**0.987**
1913 roden's corner	+	0.946
1904 the last hope	+	**0.983**
1903 barlasch of the guard	+	**0.968**
1895 the grey lady	+	**0.982**

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Table 4: The prediction results on the Lyman Frank corpus

Book	Prediction correct	Confidence (0-1)
1908 dorothy and the wizard in oz	+	**0.972**
1901 american fairy tales	+	**0.966**
1906 aunt jane’s nieces	+	**0.954**
1916 mary louise	+	0.886
1911 aunt jane’s nieces and uncle john	+	0.893
1900 the wonderful wizard of oz	+	**0.972**
1910 the emerald city of oz	+	**0.991**
1912 sky island	+	**0.986**
1902 the surprising adventures	+	**0.973**
1915 the scarecrow of oz	+	**0.979**
1907 ozma of oz	+	**0.981**
1906 aunt jane’s nieces abroad	+	**0.984**
1912 aunt jane’s nieces on vacation	+	**0.978**
1903 the enchanted island of yew	+	**0.981**

The above results were obtained from the proposed approach, which was performed on a per-book basis, based on the proposed statistical features sets. Two sets of features, Fs3 and Fs4, were mainly based on estimating probabilities from adjacent words. This kind of estimation has played a strong role in disclosing important connections between MCWs and exposing their usage patterns. As stated, the use of MCWs is a frequent practice for machine-learning authorship models. The use of statistical properties of words is related to these models, but in this case is distinct from multivariate approaches which focus on stylometry. The proportions of each other, in sequential order (e.g., the #2 MCW’s frequency proportion was divided by the #1 MCW’s frequency-proportion) forms the third feature set Fs3. This is considered as an important estimator for adjacent words’ connection; this feature was named “in-series frequency ratios” and it constitutes one of the novel contributions of this research.

Table 5: The prediction results on the Humphrey Ward corpus

Book	Prediction correct	Confidence (0-1)
1884 Miss Bretherton	+	0.978
1900 Eleanor	+	0.969
1898 Helbeck of Bannisdale 2	+	0.98
1913 The Mating of Lydia	+	0.996
1916 Lady Connie	+	0.984
1911 the case of Richard Meynell	+	0.876
1906 Fenwick’s Career	+	0.976
1888 Robert Elsmere	+	0.992
1908 The testing of Diana Mallory	+	0.981
1896 Sir George Tressady 2	+	0.996
1913 The Coryston family	+	0.979
1915 a great success	+	0.981
1914 Delia Blanchflower	+	0.993
1905 The marriage of William Ashe	+	0.984
1894 Marcella	+	0.985
1881 Milly and Olly	+	0.979
1903 Lady Rose’s Daughter	-	0.997

Table 6 has presented the overall performance of the proposed approach. It concludes many internal calculations based on the analysis of each author’s set of books. Three types of metrics were presented including sensitivity, specificity and negative likelihood ratio (LR¬-). The advantage of using the negative LR- metric was to show

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

the stability of the proposed model. This negative type of LR has been used to rule-out plagiarism. The metric showed the stability of the outcomes and reflected the reality of the results of the authorship cases. The interpretation of likelihood ratios values was based on balancing the sensitivity and specificity values.

Table 6: The overall results for all 25 classes using the proposed intrinsic plagiarism detection approach

Code	Author Name (class)	Sensitivity	Specificity	LR^-
A	Henry_Rider	0.96	0.998	0.0401
B	Kate_Douglas	0.928	1	0.0720
C	Hall_Caine	0.333	1	0.6670
D	Edith_Nesbit	1	1	0
E	Irving_Bacheller	1	1	0
F	Lyman_Frank	1	0.992	0
G	Marie_Corelli	1	0.996	0
H	Gilbert_Parker	0.941	1	0.0590
I	Henry_Seton	0.916	1	0.0840
J	Ralph_Connor	1	1	0
K	Humphrey_Ward	0.941	0.996	0.0592
L	Edith_Wharton	0.909	0.992	0.0917
M	Emerson_Hough	0.777	0.992	0.2247
N	Grant_Allen	1	1	0
O	Robert_Barr	1	1	0
P	Jerome_Kapla	0.9	0.996	0.1004
Q	Frances_Burnett	1	1	0
R	Andy_Adams	0.8	1	0.2
S	George_Augustus	0.8	0.996	0.2008
T	Stanley_John	1	1	0
U	Gertrude_Atherton	0.9	0.996	0.1004
V	Robert_Louis	0.888	1	0.1120
W	George_Gissing	1	1	0
X	Arthur_Conan	1	0.992	0
Y	Francis_Marion	0.923	0.996	0.07731
Overall results				**0.08355**

The larger the positive likelihood values the greater the indication that the text was plagiarised. While the smaller the likelihood value, the greater the indication that the text was tampered free. As stated before, a high value of specificity was more important than sensitivity in plagiarism and authorship detection, so a high specificity value indicated that the text was tampered free (it always occurs with high values of TN). This was compatible with using the LR metric, a smaller likelihood value indicated that the text was tampered-free, which means a high value for TN.

Table 7: The misclassification error (Miss-E) for each set of the proposed features based on two classification algorithms

	SVM				MLP			
F-type	FS1	FS2	FS3	FS4	FS1	FS2	FS3	FS4
Miss-E	**0.281**	**0.221**	**0.191**	**0.213**	**0.063**	**0.061**	**0.003**	**0.006**
	BN				RF			
F-type	FS1	FS2	FS3	FS4	FS1	FS2	FS3	FS4
Miss-E	**0.137**	**0.125**	**0.064**	**0.075**	**0.322**	**0.297**	**0.178**	**0.193**

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

The most informative features for the neural network based model (MLP) were those dealing with standard deviation. Especially, the features set of in-series frequency ratios of MCWs (F3 as shown in Table 7, indicated in red). This represents one of the most original contributions of this work, highlighting the importance of the relative frequencies of words as opposed to the raw frequencies.

To investigate different features' sets and classification algorithms, several authorship verification tasks were proposed. Two types of analysis procedures were applied, firstly each features set was examined separately and the misclassification error based on four classification algorithms was calculated. Table 7 presents the performance of each classifier on each proposed set of features. The third set of features has scored the lowest misclassification error (MSE) value for all classifiers algorithms. This set of features presents another original contribution of this research. Secondly, the first feature set Fs1 was examined, then the second feature set Fs2 added to Fs1 to form the second feature set (Fs1+Fs2). Fs3 set was added to form the third feature set (Fs1+Fs2+Fs3). The fourth feature set contains all four types of features (Fs1+Fs2+Fs3+Fs4). This incremental method was chosen because it represents the evolutionary sequence of style features that measures the text density [38], as shown in Table 8.

Table 8: The performance (detection accuracy) of four ML methods based on different sets of features

	SVM	MLP	BN	RF
Fs1	0.7815	**0.8823**	0.8611	0.8107
Fs1+Fs2	0.8021	**0.897**	0.8746	0.8557
Fs1+Fs2+Fs3	0.8651	**0.9396**	0.9198	0.8772
Fs1+Fs2+Fs3+Fs4 (Best)	**0.8885**	**0.9715**	**0. 9257**	**0.8625**

The two application procedures (features were analysed separately and in an accumulation way) added new types of features to the existing sets. Four classifiers were trained as classification algorithms, including support vector machines (SVM), multilayer perceptron (MLP), Bayes network (BN) and random forest (RF), respectively. Leave-one-out-cross-validation was used to estimate the accuracy of the classification model. It is obvious from Table 8 that MLP has outperformed the other algorithms by scoring 0.97 as a prediction accuracy value using all groups of features. The experimental design investigates the impact of analysing different numbers of MCWs on the performance of authorship verification. Four metrics, including specificity, sensitivity, and accuracy, were used and then the misclassification error (MSE) was calculated for each classifier, as shown in Table 9. These metrics are assumed to be the standard evaluation metrics that are used in authorship analysis and plagiarism detection [38]. The confusion matrices, giving the overall prediction performance of the proposed method, were presented in Table 9 as well.

MLP outperformed all other algorithm using the four sets of features separately as well as on using all of the features together, which confirmed the efficiency of employing jointly stylometry and MLPs. FS3 features set was the dominant set that affected the performance of all machine learning algorithms, in particular the MLP, (highlighted in red).

We further validated the results obtained by our proposed method. The proposed approach (LSA and MLP) attains the lowest misclassification error with all sets of features, compared to the Bayesian Network, Support Vector Machine and Random Forest in this order, as shown in Table 9. In order to investigate whether these results were statistically meaningful, we have conducted a t-test (with significance level $\alpha = 0.01$). The t-test was run on the z-values (standard deviations away from average) and obtained a p-value of $3.1E-6$, which shows that our proposed approach is statistically better than the compared approaches (in particular the Bayesian Network, which provides the next best performance).

ACCEPTED MANUSCRIPT

Table 9: Averaged results for four classification algorithms: MLP, BN, SVM and RF

	Multilayer Perceptron				Bayesian Networks				Support Vector Machine				Random Forest			
Misclassification error	0.0240				0.0514				0.1268				0.3802			
Accuracy	0.9715				0. 9257				0.8885				0.8625			
Specificity	0.9090				0.6818				0.0.608				0.3703			
Sensitivity	0.9857				0.9703				0.9294				0.7663			
Confusion matrix for the experiments	True	Estimated		Labels	True	Estimated		Labels	True	Estimated		Labels	True	Estimated		Labels
Labels	Labels	1	2	Totals	Labels	1	2	Totals	Labels	1	2	Totals	Labels	1	2	Totals
1: tampered	1	277	1	1	1	262	7	269	1	224	20	244	1	141	68	209
2: tampered-free	2	4	10	2	2	8	15	23	2	17	31	48	2	43	40	83
	Totals	281	11	292	Totals	270	22	292	Totals	241	51	292	Totals	241	108	292

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

6. Conclusion

The proposed intrinsic plagiarism detection method presented in this paper has relied on deriving sets of features from MCW frequencies to measure the variation in the target author writing style. Different sets of features that reflect the in-depth distribution of MCWs in each class were devised. The frequencies of particular MCWs, the relative frequencies of all MCWs, the in-series proportional frequencies (e.g., 2nd / 1st, etc.), and the z-scores were calculated. The third and fourth sets of features mainly rely on estimating the probability from adjacent words. This kind of estimation is used to disclose connections between MCWs and expose their usage patterns. A multi-layer perceptron and three other machine learning techniques were employed to generate the authorship prediction models. In order to evaluate the efficiency of these methods, a series of experiments were conducted on 25 authors' datasets from the corpus of English novels (CEN). The experimental results showed that the proposed method is able to detect the author's class when no external references collection is available. MLP outperformed the other algorithms, the most informative features were those dealing with standard deviation, especially standard deviation of the in-series proportions of MCWs. This represents one of the most significant contributions of this work, as the importance of the relative frequencies of words (as opposed to the raw frequencies) have not yet been reported in the literature. Our future work will include further testing and evaluation of the proposed approach.

References

1. Abbasi, Ahmed, and Hsinchun Chen. "Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace." *ACM Transactions on Information Systems (TOIS)* 26, no. 2 (2008): 7.

2. Alhabashneh, Obada., Iqbal, Rahat., Doctor, Faiyaz., James, Anne., "Fuzzy Rule Based Profiling Approach For Enterprise Information Seeking and Retrieval", (2017), Information Sciences, Elsevier.

3. Al Batineh, Mohammed S. "Latent Semantic Analysis, Corpus stylistics and Machine Learning Stylometry for Translational and Authorial Style Analysis: The Case of Denys Johnson-Davies' Translations into English." PhD diss., Kent State University, 2015.

4. Altınel, Berna, Murat Can Ganiz, and Banu Diri. "A corpus-based semantic kernel for text classification by using meaning values of terms." *Engineering Applications of Artificial Intelligence* 43 (2015): 54-66.

5. Alzahrani, Salha M., Naomie Salim, and Ajith Abraham. "Understanding plagiarism linguistic patterns, textual features, and detection methods." *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 42, no. 2 (2012): 133-149.

6. Biber, Douglas. *Dimensions of register variation: A cross-linguistic comparison*. Cambridge University Press, 1995.

7. Boukhaled, Mohamed Amine, and Jean-Gabriel Ganascia. "Using Function Words for Authorship Attribution: Bag-Of-Words vs. Sequential Rules." *Natural Language Processing and Cognitive Science: Proceedings 2014* (2015): 115.

8. Burrows, John F. "Word-patterns and story-shapes: The statistical analysis of narrative style." *Literary and linguistic Computing* 2, no. 2 (1987): 61-70.

9. Burrows, John. "'Delta': A measure of stylistic difference and a guide to likely authorship." *Literary and linguistic computing* 17, no. 3 (2002): 267-287.

10. Ceska, Zdenek. "Automatic plagiarism detection based on latent semantic analysis." PhD diss., Ph. D. dissertation, Faculty Appl Sci., Univ. West Bohemia, Pilsen, Czech Republic, 2009.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

11. Cosma, Georgina, and Mike Joy. "An approach to source-code plagiarism detection and investigation using latent semantic analysis." *IEEE transactions on computers* 61, no. 3 (2012): 379-394.

12. Craig, Hugh. *Stylistic analysis and authorship studies*. Blackwell Publishing, 2004.

13. Diederich, Joachim, Jörg Kindermann, Edda Leopold, and Gerhard Paass. "Authorship attribution with support vector machines." *Applied intelligence* 19, no. 1-2 (2003): 109-123.

14. Edmunds, Angela, and Anne Morris. "The problem of information overload in business organisations: a review of the literature." *International journal of information management* 20, no. 1 (2000): 17-28.

15. Fucks, Wilhelm. "On mathematical analysis of style." *Biometrika* 39, no. 1/2 (1952): 122-129.

16. Juola, Patrick, and R. Harald Baayen. "A controlled-corpus experiment in authorship identification by cross-entropy." *Literary and Linguistic Computing* 20, no. Suppl (2005): 59-67.

17. Kakkonen, Tuomo, and Maxim Mozgovoy. "Hermetic and web plagiarism detection systems for student essays—an evaluation of the state-of-the-art." *Journal of Educational Computing Research* 42, no. 2 (2010): 135-159.

18. Koch, Mark W., Mary M. Moya, Larry D. Hostetler, and R. Joseph Fogler. "Cueing, feature discovery, and one-class learning for synthetic aperture radar automatic target recognition." *Neural Networks* 8, no. 7 (1995): 1081-1102.

19. Koppel, Moshe, and Jonathan Schler. "Authorship verification as a one-class classification problem." In *Proceedings of the twenty-first international conference on Machine learning*, p. 62. ACM, 2004.

20. Koppel, Moshe, Jonathan Schler, and Kfir Zigdon. "Determining an author's native language by mining a text for errors." In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pp. 624-628. ACM, 2005.

21. Koppel, Moshe, Jonathan Schler, and Shlomo Argamon. "Computational methods in authorship attribution." *Journal of the American Society for information Science and Technology* 60, no. 1 (2009): 9-26.

22. Landauer, Thomas K., Peter W. Foltz, and Darrell Laham. "An introduction to latent semantic analysis." *Discourse processes* 25, no. 2-3 (1998): 259-284.

23. Luyckx, Kim, and Walter Daelemans. "Authorship attribution and verification with many authors and limited data." In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pp. 513-520. Association for Computational Linguistics, 2008.

24. Mendenhall, Thomas Corwin. "The characteristic curves of composition." *Science* (1887): 237-249.

25. Mosteller, Frederick, and David Wallace. "Inference and disputed authorship: The Federalist." (1964).

26. Merriam, Thomas VN, and Robert AJ Matthews. "Neural computation in stylometry II: An application to the works of Shakespeare and Marlowe." *Literary and Linguistic Computing* 9, no. 1 (1994): 1-6.

27. Oakes, Michael P., and Meng Ji, eds. *Quantitative methods in corpus-based Translation Studies: A practical guide to descriptive translation research*. Vol. 51. John Benjamins Publishing, 2012.

28. Roberts, Stephen, and Lionel Tarassenko. "A probabilistic resource allocating network for novelty detection." *Neural Computation* 6, no. 2 (1994): 270-284.

29. Salton, Gerard, Anita Wong, and Chung-Shu Yang. "A vector space model for automatic indexing." *Communications of the ACM* 18, no. 11 (1975): 613-620.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

30. Sebastiani, F., 2002. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, *34*(1), pp.1-47.

31. Smith, Tony C., and Ian H. Witten. "Language inference from function words." (1993).

32. Stamatatos, Efstathios. "A survey of modern authorship attribution methods." *Journal of the American Society for information Science and Technology* 60, no. 3 (2009): 538-556.

33. Tax, David MJ, and Robert PW Duin. "Uniform object generation for optimizing one-class classifiers." *Journal of Machine Learning Research* 2, no. Dec (2001): 155-173.

34. Tweedie, Fiona J., Sameer Singh, and David I. Holmes. "Neural network applications in stylometry: The Federalist Papers." *Computers and the Humanities* 30, no. 1 (1996): 1-10.

35. Yule, G. Udny. "On sentence-length as a statistical characteristic of style in prose: With application to two cases of disputed authorship." *Biometrika* 30, no. 3/4 (1939): 363-390.

36. Zechner, Mario, Markus Muhr, Roman Kern, and Michael Granitzer. "External and intrinsic plagiarism detection using vector space models." In *Proc. SEPLN*, vol. 32, pp. 47-55. 2009.
37. Zhao, Ying, and Justin Zobel. "Searching with style: authorship attribution in classic literature." In *Proceedings of the thirtieth Australasian conference on Computer science-Volume 62*, pp. 59-68. Australian Computer Society, Inc., 2007.

38. Zheng, Rong, Jiexun Li, Hsinchun Chen, and Zan Huang. "A framework for authorship identification of online messages: Writing-style features and classification techniques." *Journal of the American Society for Information Science and Technology* 57, no. 3 (2006): 378-393.

39. Zu Eissen, Sven Meyer, Benno Stein, and Marion Kulig. "Plagiarism detection without reference collections." In *Advances in data analysis*, pp. 359-366. Springer Berlin Heidelberg, 2007.

40. Zurini, Madalina. "Stylometry Metrics Selection for Creating a Model for Evaluating the Writing Style of Authors According to Their Cultural Orientation." *Informatica Economica* 19, no. 3 (2015): 107.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

An Integrated Approach for Intrinsic Plagiarism Detection

Biography

Muna Al-Sallal

Photo will be provided.

Muna Al-Sallal is a final year PhD candidate in the Faculty of Engineering, Environment and Computing at Coventry University. She investigates a number of machine learning approaches for plagiarism detection. With the rapid proliferation of Internet technologies and applications, misuse of online materials for inappropriate purposes has become a major concern for the integrity of science and societies. She has participated in a number of international conferences and workshops.

Dr Rahat Iqbal

Dr Rahat Iqbal is Managing Director of Interactive Coventry Ltd and a Reader/Associate Professor in the Faculty of Engineering, Environment and Computing at Coventry University. He has a track record of project management and leadership of industrial projects funded by EPSRC, TSB, ERDF and local industries (e.g. Jaguar Land Rover Ltd, Trinity Expert Systems Ltd). He was involved in the project management and development of the EU FP7 project CHIL (Computers in Human Interaction Loop) at the Technical University of Eindhoven , Netherlands. Recently, he has successfully led a project in collaboration with Jaguar Land Rover on self-learning car for predicting driver's behaviour for personalisation of telematics and optimisation of route planning. He has managed many industrial projects, in Intelligent Systems, Predictive Modelling, User Behaviour, Information Retrieval and Fault Detection. He has published more than 100 papers in peer-reviewed journals and reputable conferences and workshops. Dr Iqbal is on the programme committee of several international conferences and workshops. He is also a fellow of the UK Higher Education Academy (HEA). Dr Iqbal has also edited several special issues of international journals within the field of Information Retrieval and User Supportive systems.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Dr Vasile Palade

Dr. Vasile Palade is a Reader in Pervasive Computing in the Faculty of Engineering and Computing and a member of the Cogent Computing Applied Research Centre at Coventry University. He previously had academic and research positions at the University of Oxford - UK (Departmental Lecturer in the Department of Computer Science), University of Hull - UK (Research Fellow in the Department of Engineering) and the University of Galati - Romania (Associate Professor in the Department of Computer Science and Engineering).

His research interests lie in the area of machine learning/computational intelligence, and encompass mainly neuro-fuzzy systems, various nature inspired algorithms such as swarm optimization algorithms, hybrid intelligent systems, ensemble of classifiers, class imbalance learning. Application areas include Bioinformatics problems, fault diagnosis, web usage mining, among others.

Dr. Palade is author and co-author of more than 100 papers in journals and conference proceedings as well as books on computational intelligence and applications. He has also co-edited several books including conference proceedings. He is an Associate Editor for several journals, such as *Knowledge and Information Systems* (Elsevier), *International Journal on Artificial Intelligence Tools* (World Scientific), *International Journal of Hybrid Intelligent Systems* (IOS Press), *Neurocomputing* (Elsevier). He has delivered keynote talks to international conferences on machine learning and applications

Dr. Saad Amin

Saad Amin is the Post-graduate Programme Manager and Principal Lecturer in Network Computing in the Faculty of Engineering and Computing.His main area of research is Health Informatics.

One of his researches interests centres on the design of biomedical image processing algorithms and multimedia applications on parallel computers and distributed computing, and their implementation on

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Cluster of workstations (CoW). He is also involved in the area of context-aware multimedia information systems for e-health decision support.Previously, he was involved with a European Community project at the Parallel Algorithms Research Centre, Loughborough University.

The main objectives of the project were to establish a European standard platform for the exploration of biocomputational, nanotechnological and holographic algorithms. Dr Amin is the principal investigator of several funded projects. He has supervised eight completed PhD students and has eight current PhD students. He has published more than 70 papers in peer-reviewed journals, reputable conferences and books chapters on many aspects of this research. He has organised international conferences and workshops, as well as acted as Session Chair and on program committees for many international conferences.

He served on the IEEE Committee for Signal Processing Chapter for the UK and Ireland and Vice-Chairman of the British Computer Society/Middle East section.

Dr Victor Chang

Victor Chang is an Associate Professor (Reader) of Suzhou Business School, Xi'an Jiaotong Liverpool University, China. He was previously a Senior Lecturer in the School of Computing, Creative Technologies at Leeds Beckett University, UK. He’s a Visiting Researcher at the University of Southampton, UK and an Honorary Associate Professor at the University of Liverpool, UK. He is an expert on Cloud Computing and Big Data in both academia and industry with extensive experience in related areas since 1998. He completed a PGCert (Higher Education) and PhD (Computer Science) within four years while working full-time. He has over 100 peer-reviewed published papers. He won £20,000 funding in 2001 and £81,000 funding in 2009. He was involved in part of the £6.5 million project in 2004, part of the £5.6 million project in 2006 and part of a £300,000 project in 2013. He

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

won a 2011 European Identity Award in Cloud Migration, since his work is making contributions. He has won 2016 European Identity and Cloud Award on the best project in research, involved with more than 20 collaborators worth more than $10 millions in valuation. He was selected to present his research in the House of Commons in 2011 and won the best paper in 2012 and 2015. He has demonstrated Storage as a Service, Health Informatics as a Service, Financial Software as a Service, Education as a Service, Big Data Processing as a Service, Integration as a Service, Security as a Service, Social Network as a Service, Data Visualization as a Service (Weather Science) and Consulting as Service in Cloud Computing and Big Data services in both of his practitioner and academic experience. His proposed frameworks have been adopted by several organizations. He is the founding chair of international workshops in Emerging Software as a Service and Analytics and Enterprise Security. He is the founding chair of IoTBDS and COMPLEXIS which have become popular in research communities. He is an Editor-in-Chief (EIC) in International Journal of Organizational and Collective Intelligence and a founding EIC in Open Journal of Big Data. He is the Editor of a highly prestigious journal, Future Generation Computer Systems (FGCS). He is a reviewer of numerous well-known journals and had published three books on Cloud Computing which are available on Amazon website. He is a keynote speaker for CLOSER 2015/WEBIST2015/ICTforAgeingWell 2015 and has received positive support.

ACCEPTED MANUSCRIPT

*Highlights (for review)

ACCEPTED MANUSCRIPT

An Integrated Approach for Intrinsic Plagiarism Detection

Highlights

- We present an approach for intrinsic plagiarism detection for the next generation web.
- The approach is based on Stylometry, LSA and MLP.
- Stylometry investigates the latent changes in writing style,
- LSA analyses words co-occurrence and most common words
- MLP is used to develop the classifier model of authorship detection.
- The proposed approach has outperformed the existing methods.

ACCEPTED MANUSCRIPT