


# Arabic Authorship Attribution Using Synthetic Minority Over-Sampling Technique and Principal Components Analysis for Imbalanced Documents

Hassina Hadjadj, USTHB University, Algeria

 <https://orcid.org/0000-0002-5409-6576>

Halim Sayoud, USTHB University, Algeria

## ABSTRACT

Dealing with imbalanced data represents a great challenge in data mining as well as in machine learning task. In this investigation, the authors are interested in the problem of class imbalance in authorship attribution (AA) task, with specific application on Arabic text data. This article proposes a new hybrid approach based on principal components analysis (PCA) and synthetic minority over-sampling technique (SMOTE), which considerably improve the performances of authorship attribution on imbalanced data. The used dataset contains seven Arabic books written by seven different scholars, which are segmented into text segments of the same size, with an average length of 2,900 words per text. The obtained results of the experiments show that the proposed approach using the SMO-SVM classifier presents high performance in terms of authorship attribution accuracy (100%), especially with starting character-bigrams. In addition, the proposed method appears quite interesting by improving the AA performances in imbalanced datasets, mainly with function words.

## KEYWORDS

Arabic Language, Authorship Attribution, BayesNet, Imbalanced Datasets, Principal Component Analysis (PCA), SMO-SVM, Synthetic Minority Over-Sampling Technique (SMOTE)

## 1. INTRODUCTION

Authorship attribution (AA) is one of the earliest research fields of computational linguistics and has a long history in identifying disputed or unknown authors (Mosteller & Wallace, 1984). Several researchers were interested in a myriad of applications of AA such as email authorship verification, categorizing harassing emails and anonymous messages in textual conversations and social media forensics (Rocha et al., 2017), Online criminality (Edwards, 2018). In addition, AA can be used to identify the document sources (Li et al., 2013), disputed authorship (Eder, 2015), plagiarism detection in student essays (AlSallal et al., 2019), etc.

AA consists of studying the author's writing pattern (or stylometry) to respond to the following question: Who is the author of this document?. Accordingly, the suitable set of features is extracted and combined with the more reliable classification technique to find the right author. In this regard, function words (stop words) and the spelling errors should be kept, because they have a substantial

DOI: 10.4018/IJCINI.20211001.0a33

This article published as an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>) which permits unrestricted use, distribution, and production in any medium, provided the author of the original work and original publication source are properly credited.

role in the identification task. On the other hand, two parameters are important in stylometry and should be exploited, namely: the text's length (number of words) and the number of authors. In addition, some researchers have set some conditions to accurately identify the authors, such as the same theme, the same genre (i.e. poems, news, scientific papers, etc.) and the same period of time. However, features extraction is not the only operator that influences AA, where there are other factors such as the dataset size (training and test), number of candidate authors and the distribution of the training corpus over the authors (balanced or unbalanced dataset).

Over decades, several stylometric features have been investigated and applied in AA, where there are a myriad of stylometric features commonly used in stylometry such as sentence length and vocabulary richness (Yule, 1994), function words (Holmes et al., 2001; Zhao & Zobel, 2005), punctuation marks (Baayen et al., 2002) and characters n-gram (Juola, 2004). The use of function words to produce best performances is due to two reasons. The first one is their frequency in the document, which is very hardly under conscious control and reduces the risk of false attribution. Secondly, function words, unlike content words, are totally independent from the text's topic or genre (Argamon et al., 2007).

Regarding the number of related works carried out in AA, most of them addressed Latin languages (e.g. English), while a few researches were conducted on the Arabic language and in particular those using unbalanced data. Several researchers confirmed that balanced datasets provide high accuracies in contrast to unbalanced datasets (Li et al., 2018; Pan et al., 2020). However, it is tricky to collect sufficient data for each author. In this regard, the aim of this investigation is to deal with Arabic AA using unbalanced dataset, where seven Arabic books with different text lengths and written by different authors (in the same period) have been used.

To resolve AA problem, we have used the unsupervised principal component analysis (PCA) combined with the oversampling technique (SMOTE) to reduce data dimensionality. We have conducted a series of experiments using our dataset (i.e. SAB-2 dataset), where PCA was applied to eliminate irrelevant features, and subsequently SMOTE resampling was used to balance the class distribution and increase the variety of sample domains. Finally, SMO-SVM and BayesNet classifiers were applied on the filtered dataset, where they were compared using different evaluation metrics. The hybrid approach combining both algorithms showed interesting performances (100% of accuracy) on unbalanced data.

This paper is organized as follows: in section 2, we present some related works on AA and Arabic AA. The dataset is described in section 3, while section 4 presents our AA approach. Finally, we present the experimental results in section 5, and section 6 gives a short conclusion on this research work.

## 2. RELATED WORKS

In the last decade, some of AA researches have achieved high accuracies using lexical features such as the frequency of common words, stop words, words n-gram and characters n-gram (Argamon & Levitan, 2005). As mentioned above, a lot of works addressed English texts, and some authors used the words. For instance, Abbasi and Chen (2005) used a set of 150 function words, Zaho and Zobel (2005) used a set of 365 function words. Similarity, another set of 645 function words was proposed in (Argamon et al., 2007).

On the other hand, Koppel et al. (2007) used the 250 most frequent words, while Stamatatos extracted the 1000 most frequent words (Stamatatos, 2006a). On large scale, Madigan et al. (2005) used all the words that appear at least twice in the dataset. Conversely, Koppel and Schler (2003) suggested various writing mistake measures to discover the idiosyncrasies of an author's style. Thus, a set of misspelling errors was defined, which could be used in several NLP applications to profile the writing style (Grieve, 2007).

The application of characters n-gram is the most successful approach on authorship attribution (Koppel et al., 2011), but the reason for their success is not well understood. Juola (2004) proposed

one of the best performing approaches based on characters n-gram in AA. The latter produced high performances and used in several works (Peng et al., 2003; Kešelj et al., 2003; Stamatatos, 2006b). Moreover, Grieve (2007) compared different lexical and character features, and the results showed that characters n-gram are more reliable. Finally, NN (neural network) classifier based on continuous representation of n-gram features was used, and the results showed that the proposed model outperforms the state-of-the-art on two datasets (Sari et al., 2017).

To deal with text classification, a first step is to get a representation of the text for using in machine learning. Term Frequency-Inverse Document Frequency (TF-IDF) is a common way used to convert text into numerical dataset (Salton & Buckley, 1988). Although this method has been proved in Information Retrieval field and text mining task, but it is not most efficient for authorship attribution because TFIDF disregard the class label information of the training texts (Swathi et al., 2018).

The choice of the suitable classifier is an important step in classification problems. Hence, Multiple Regression and Discriminant Analysis were used by Stamatatos et al. (2001) in AA, and the best accuracy was 72%. Kjell (1994) investigated neural networks and bayesian classifiers in AA, and the best achieved accuracy was 90%. Bozkurt et al. (2007) compared several classification techniques and features such as bayesian classifier, parametric method, Parzen windows, histogram based method, k-nearest-neighbor, SVM, k-means clustering algorithm and NN. In addition, they also used PCA in conjunction with various classifiers to reduce the features dimension. The experimental results showed that SVM combined with PCA and based on function words reported the best performances. Various AA models make use of SVM classifiers with some lexical or syntactic feature. Some studies comparing different machine learning techniques in AA (Abbasi & Chen, 2005; Zhang & Lee, 2006) showed that SVM is the best learning model for AA. Other studies figured out that some variations of Winnow and Bayesian regression are also very favorable (Koppel & Schler, 2003; Genkin et al., 2007).

Another important point in AA is to find the minimal amount of data in order to get reliable results. In 2015, Eder (2015) tried to resolve this problem by providing some key solutions to find the minimal size of text samples in AA. The experiments were conducted on different types of texts and different languages, where the results showed that texts with 2500 tokens at least accurate the identification task. In addition, a novel approach has been proposed based on artificial neural network (ANN) and feature selection using the principal component analysis (PCA) technique for PDF malware detection. The evaluation on 105,000 real-world PDF documents shows that the model with PCA can significantly minimize learning time and feature redundancy with lower effect on data information loss (Zhang, 2019).

Despite the most of works in AA addressed different languages, there are a few works carried out on the Arabic language. For instance, Alwajeeh et al. (2014) used SVM and NB classifiers to deal with Arabic AA, and the results showed that both algorithms perform better. The use of words function was investigated in Arabic AA by Shaker and Corne (2012), where it was compared it to English AA. Moreover, he also proposed a hybrid approach between evolutionary search and LDA to get the best small number of features. Ouamour and Sayoud (2018) presented a method to handle AA problem of short Arabic texts. This survey based on several features such as characters, characters bi-gram, characters tri-gram, characters tetra-gram, words, words bi-gram and rare words. The AA is achieved by 3 classifiers and a new proposed fusion called VBF (Vote Based Fusion). The results were interesting and the proposed VBF fusion produced high accuracy (about 90%). Furthermore, Authorship attribution in Arabic poetry Model (APAAM) is proposed by Al-Falahi et al. (2019). This study based on different features such as poetry features, syntactic features and semantic features using classification algorithms Linear discriminant analysis, Naïve Bayes and Support Vector Machine. The experiments were conducted on a set of 114 random poets from entirely various periods. The results were interesting with 99.12% of performance accuracy.

Unbalanced data is the common issue in data mining and machine learning, and it was solved mainly in two ways such as data level and algorithm level (Krawczyk, 2016). On data level, the resampling technique is the most common technique applied to get balanced distribution from

unbalanced datasets. Different resampling techniques have been proposed, and can be categorized into two groups, i.e. undersampling and oversampling. Undersampling removes certain numbers of instances from the original dataset by randomly select a set of major class examples, and subsequently remove this sample to achieve a balanced dataset. On the other hand, oversampling is one of the simplest sampling methods, and consists of randomly selecting a set of minor class examples (He & Garcia, 2009), and subsequently duplicating the selected examples and increasing them to original data set. Moreover, one of the most famous approaches in unbalanced class problem is the Synthetic Minority Oversampling Technique (or SMOTE) (Chawla et al., 2002). SMOTE can be better than simple oversampling and it has proven successful in wide variety of applications from different fields (He et al., 2009). For example, SMOTE was used for diagnose the disease's class, where it is used in (Mohd et al., 2019) to balanced training dataset by generating new synthetic samples, where the balanced datasets were trained with machine learning algorithms. The experiment is conducted on oral cancer dataset and erythemato-squamous diseases dataset from the UCI machine learning datasets. SMOTE approach showed best results in clinical disease classification. In addition, two effective sampling methods based on SMOTE and Gaussian distribution are presented, where the first one, Adaptive-SMOTE, improves the SMOTE method and the other technique, Gaussian Oversampling, combines Gaussian distribution with the dimension reduction (Pan et al., 2020). The experiments on 15 datasets display that the two proposed sampling methods better compared with other typical methods. Furthermore, Soltanzadeh and Hashemzadeh (2021) improved SMOTE-based algorithm, namely Range-Controlled SMOTE (RCSMOTE), which aimed all three problem issues of SMOTE approach together. This method present advances the over-sampling process on the right samples and in the right places in the data space. The experiments on several datasets, display that the RCSMOTE overcomes the problems of SMOTE. The study of (Alhakbani, 2018) focuses on the class imbalance problem in data mining, where swarm intelligence techniques like Stochastic Diffusion Search (SDS) and Dispersive Flies Optimization (DFO) along with SVM classifier were used. Results display that SDS can be improved the classifier performance on imbalanced datasets. In addition, DFO has given promising results in these experiments; moreover, Stamatatos (2008) try to cover the class unbalance problem for authorship attribution task. In brief, a new method for handling unbalanced datasets through segmenting the training texts into samples was proposed. Hence, Majority class can be segmented into longer samples, and minority class into many short samples. Thus, the training size of minority class is increased by text sampling. This work is experimented on Arabic and English text corpora.

In this study, we propose a hybrid method based on PCA and SMOTE approach to enhance the performances of AA in unbalanced Arabic dataset.

### **3. DATASET**

We have created a new dataset for Arabic authorship attribution, where it regroupes seven different Arabic books written by seven religious scholars. The dataset is called SAB-2 (Seven Arabic Books – dataset two). The books are clustered into distinct text segments, where the segment lengths are not equal and range within an interval. The average length is 2900 words per segment. In fact, according to the previous research of Eder (2015), it has been shown that the minimum size per text should be at least 2500 words to ensure a good authorship performance. The size of each segment per book is shown in Table 1.

### **4. AUTHORSHIP ATTRIBUTION BASED ON PCA AND SMOTE**

The main purpose of this work is to increase the identification performance in the case of unbalanced dataset. Hence, we propose a new hybrid approach based on principal components analysis (PCA) and synthetic minority over-sampling technique (SMOTE) with a new set of features. Firstly, the

**Table 1. SAB-2 dataset description. Big and small are logical parameters (binary value)**

Book/Author	Number of segments by book	Big/ Small parameter#
1st book: books of Hassan	29 segments	Big
2nd book: books of alarifi	8 segments	Small
3rd book: books of Alghazali	39 segments	Big
4th book: books of AlQuaradhawi	13 segments	Small
5th book: books of Abdelkafy	10 segments	Small
6th book: books of Aid Alkarny	23 segments	Big
7th book: books of Amrokhaled	9 segments	Small

frequencies of the lexical based stylometric features are calculated and reduced by PCA, and next the data is balanced using SMOTE. The proposed AA method used in this work, is summarized in the following block diagram (Figure 1).

#### 4.1. Text Pre-Processing

In order to improve data quality and the identification performance, the textual data must be pre-processed. In this regard, punctuation marks, diacritics, numbers and non-Arabic letters are removed from the texts. Next, each text is encoded according to UTF8 encoding.

#### 4.2. Feature Extraction

The choice of the best set of features, which can characterize the author's writing style, is a primordial step for machine learning algorithms, and is highly related to the text language. Several linguistic features are proposed in AA. For instance, vocabulary based features (Juola, 2006) such as the average sentence length, and syntax based features (Stamatatos et al., 2001) such as function words and characters based features. In this paper, two types of features are proposed, i.e. function words (FW) and starting n-grams.

##### 4.2.1. Function Words

Function words have little lexical content and repeated frequently in any text such as definite and indefinite articles, conjunctions, adverbs, etc. Indeed, the author's writing style could be distinguished regarding the function words. The reason behind function words to perform well is due to topic-independent (Argamon et al., 2007). In this work, we have proposed a new list of Arabic function words regrouping 600 words (Table 2).

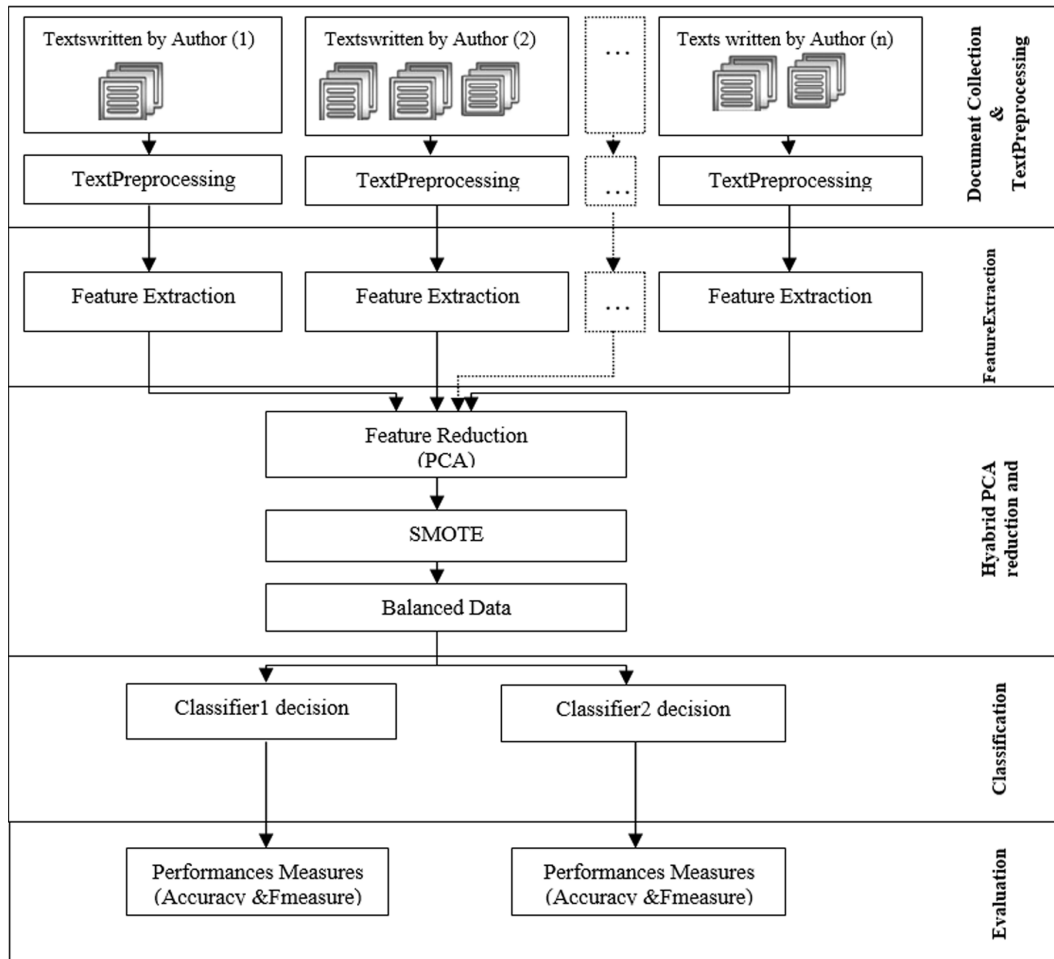
##### 4.2.2. Starting N-Grams

In order to extract this type of features, we firstly extract a list of words from the text, and subsequently we take the first characters n-gram of each word. Consecutively, based on the characters n-gram a profile is created containing the frequency of each n-gram (number of occurrences). In this work, we have used characters bi-gram and tri-gram (Figure 2).

#### 4.3. Dimensionality Reduction (Principal Components Analysis)

Dimensionality reduction is process of reducing the number of feature set in order to get a set of uncorrelated principal features. It consists of transforming the data from high-dimensional space to low-dimensional space (Guandong et al., 2013). In order to reduce the number of the original features and select the most relevant ones, PCA dimensionality reduction technique is used in this work.

Figure 1. General scheme of the combined PCA with SMOTE



PCA is a multivariate statistical method that attempts to reduce very large data set by eliminating the original features dependent on the correlation between features (Jolliffe, 2002). This elimination is achieved by a linear vector transformation of features vector. PCA aims to find maximum variance of projection of the original feature space in new subspace with fewer dimensions. This issue can be solved by determining a corresponding eigenvectors with the maximum eigenvalues in the transformed

Table 2. List of some Arabic function words

Demonstrative pronouns	Translation	Possessive pronouns	Translation	conjunctions	Translation
أنا	I	لي	Mine	لهذا	For this
نحن	We	لنا	Our	بسبب	Because of
انت	You	لك	Yours	فسوف	It will
هو	He	له	His	فلذلك	So it is

Figure 2. Example of extracting the Starting n-grams from the word of a text

ليس اليتيم من مات والده إن اليتيم يتيم العلم و الأدب	<i>Text</i>
It is not an orphan whose father dies; an orphan is an orphan of knowledge and literature	<i>Translation</i>
[ليس][اليتيم][من][مات][والده][إن][اليتيم][يتيم][العلم][و][الأدب]	<i>List of Word</i>
[لى][ال][من][ما][وا][إن][ال][يت][ال][ال]	<i>Starting Bigrams</i>
[ليس][الى][مات][وال][الى][يتي][الع][الأ]	<i>Starting Trigrams</i>

space variance matrix. Then, belonging uncorrelated component can be selected beginning by higher ones. Features with high linear correlation that have the same information are eliminated.

#### 4.4. Synthetic Minority Oversampling Technique (SOMTE) Approach

Recently, the scenario of class unbalance becomes a great challenge for data classification. There are different methods available for class unbalance problem, one of the famous approaches to solve class unbalance problem is sampling. The goal of sampling methods is to modify the distributions of the majority and minority class in the training data set to get relatively balanced class distribution (Hoang et al., 2009). The main approaches to deal with class unbalance attempt to rebalance the training set by: Under-sampling of the majority class and Over-sampling of the minority class.

One of the famous oversampling approaches is SOMTE (Synthetic Minority Over-sampling TEchnique). SMOTE was introduced by Chawla et al. (2002). The SMOTE algorithm creates artificial new instances based on the feature space similarities between existing minority examples. New instance values are derived from interpolation rather than extrapolation, so they still carry relevance to the underlying dataset. Specifically, for each minority class instance SMOTE interpolates values using a k-nearest neighbor technique and creates attribute values for new data instances. For a given integer K, the k-nearest neighbors are defined as the K elements of a set of minority class samples whose Euclidian distance exhibits the smallest magnitude along the n-dimension of feature space. To create a synthetic sample, randomly select one of the k-nearest neighbors, then multiply the corresponding feature vector difference with random number between [0,1], and add this vector to the instance as shown in Equation 1.

$$Y_{new} = Y_i + (Y'_i - Y_i) \times \partial \quad (1)$$

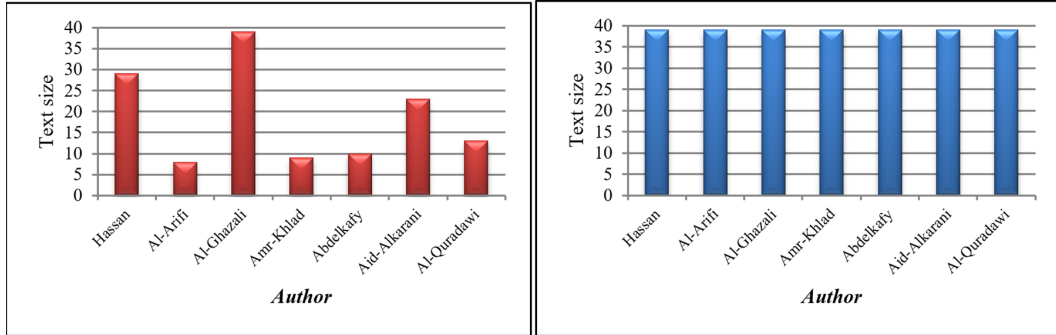
Where  $Y_{new}$  represents a new instance,  $Y_i$  is the minority under consideration  $Y'_i$  is one of the k-nearest neighbors for  $Y_i$  and is a random value between 0 and 1.

SMOTE is considered as one of the most influential data sampling or preprocessing algorithms in machine learning and data mining (García et al., 2016). Due to its popularity and influence, we have used the SMOTE approach in our work. Figure 3 shows examples of producing artificial unbalanced distributions of the data set over 7 authors before and after applying the SMOTE approach.

#### 4.5. Classification Methods

In our experiments, the authorship is classified using two classifiers and 5-fold cross validation which divides the SAB-2 dataset into training and test data. The description of the two classifiers is as follows:

**Figure 3. Distribution of text samples over the authors. Left: original imbalanced text samples. Right: Balanced training text samples produced by SMOTE approach**



#### 4.5.1. Sequential Minimal Optimization Based Support Vector Machine (SMO-SVM)

A Support Vector Machine (SVM) is a discriminative classifier, and represents the samples as points in the space. The samples are separated as categories by a clear gap as wide as possible. The clear gap (called margin) maximizes the distance between itself and nearest training point that can be used in classification, regression, clustering or other tasks. The nearest data point to the margin is known as support vector. Concerning the Sequential Minimal Optimization (SMO) algorithm, it is used to speed up the training of the SVM (Keerthi et al., 2001).

#### 4.5.2. Bayes Net (Bayesian Network)

The Bayesian network is mainly based on Bayes theorem. Its structure is an acyclic directed graph (Heckerman et al., 1995) for estimating probabilistic relationships based on conditional probabilities, where the conditional probability on each node is calculated and formed. Then, the nodes, links, and probability distributions are the structure of the network. There are two learning steps in Bayesian networks, network learning and learning the probability tables. The network structure is specified by identifying which attributes have the strongest dependencies between them. Every node  $a_i$  has a rearward probability distribution derived from its parents. The attribution includes the computation of the joint probability of different attributes, taking dependencies into account illustrated in Equation 2:

$$P(a_1, \dots, a_n) = \prod_i P(a_i / \text{Parents}(a_i)) \quad (2)$$

## 5. EXPERIMENTAL RESULTS AND ANALYSIS

### 5.1. Evaluation

Evaluation metrics play an important role to evaluate the classification performance. Accuracy measure and F-measure are one of the most relevant measures for unbalanced data. F-measure is defined as the harmonic mean of recall and precision (Hoang et al., 2009). The formula is described as follows:

$$F - \text{measure} = \frac{2 \times \text{Recall} \times \text{precision}}{\text{Recall} + \text{precision}} \quad (3)$$



**Table 3. Performance results with five-fold cross validation using SVM and Bayes Net classifiers**

Classification Algorithm	Accuracy%			F-measure		
	FW	Starting bigrams	Starting trigrams	FW	Starting bigrams	Starting trigrams
SMO-SVM	93.93	95.45	95.45	0.936	0.954	0.954
Bayes Net	90.90	93.93	92.42	0.910	0.939	0.923

Accuracy is calculated, in our investigation, by using the following formula:

$$\text{Accuracy} = \frac{\text{Number of correctly classified segments}}{\text{Total number of tested examples}} \quad (4)$$

When the classification problem has a small dataset, it would be difficult to supply sufficient data for separating the training set and testing set. So in this case, it is possible to use n-fold cross validation technique (Weiss & Kulikowski, 1991). The latter is a statistical technique generally applied in Machine Learning models, to evaluate in a significant manner the algorithms by using the overall dataset (both training and testing). In our work, we performed 5-fold cross validation. In brief, our data was randomly splitted into five groups (of equal size). The authorship model is trained on four partitions, and tested on the remaining one. The procedure is then repeated and each fold is held out for testing. Therefore, the classification task is performed 5 times, each time different partition is used as testing data, and the remaining four partitions are used in training. The results of these five classification tasks are then combined for calculating the average results for the dataset.

## 5.2. Classification Experiments Without Preprocessing

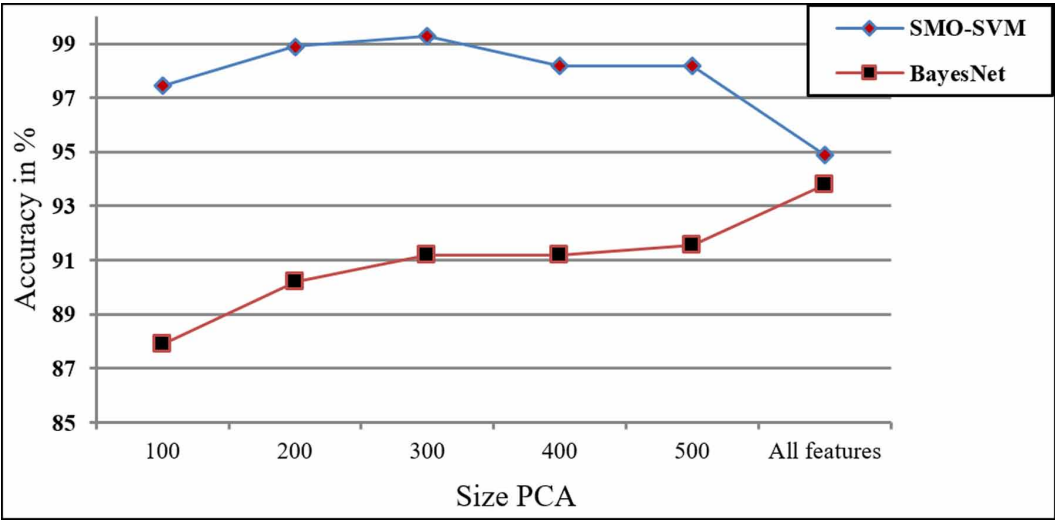
In order to figure out the robustness of our proposal, we conduct an experiment on AA without preprocessing. Thus, we use this experiment as a baseline compare different scenarios. The results produced by the two classifiers, i.e. SVM and NB, in terms of accuracy and F-measure are described in Table 3. From the table, we notice that the average accuracy ranges between 91% and 95%, and it could be considered as good. In addition, it is noticed that SVM with starting bi-grams produces good performances in contrast to other classifiers with different features.

## 5.3. Finding The Optimal PCA Dimension

In this experiment, we aim to obtain an optimal PCA dimension that represents the number of features providing the best performance. Some results of this experiment are shown in Figures 4 and 5. The performance of our method using FW by varying the number of features proposed by PCA is shown in figure 4. With both SVM and NB classifiers, we notice that the optimal size of features for FW is between 300 and 400, which represents approximately the half of the total size of the original features (i.e. 600).

Figure 5 shows the performance of our method using starting bigrams and starting trigrams by varying the number of features proposed by PCA. For starting bigrams, we notice that the optimal size of features is between 300 and 400, which represents one third of the total size of the original features (i.e. 890). We notice that the optimal size of features for starting trigrams is about 300, and represents one third of the total size of the original features (i.e. 1000). In other words, the PCA reduces the feature size to a one third. We notice that our approach begins to be accurate with 300 features using SVM, while NB classifier gets high performance with all the features.

Figure 4. Comparison between different numbers of features proposed by PCA. The case of function words



The eigenvalue latencies for the all features of the data are shown in Figures 6 and Figure 7. These figures show two curves, the upper shows the cumulative variance obtained by the all components, while the lower shows the proportion of variance for each principal component. It can be seen that about 99% of the variances (information) contained in our data are retained by the 150 principal components.

5.4. Author Identification Using SMOTE and PCA+SMOTE

In order to solve the problem of unbalanced data, we applied the PCA reduction and the SMOTE approach. We aim to obtain a new dataset with a balanced number of instances in each class. Then, we compare the obtained results with experimental results without preprocessing.

The results produced by the SMOTE and our method (SMOTE+PCA) using the three features (FW, Starting Bigrams and Starting Trigrams) are shown in table 4.

Figure 5. Comparison between different numbers of features proposed by PCA. Left: The case of starting bigrams. Right: The case of starting trigrams.

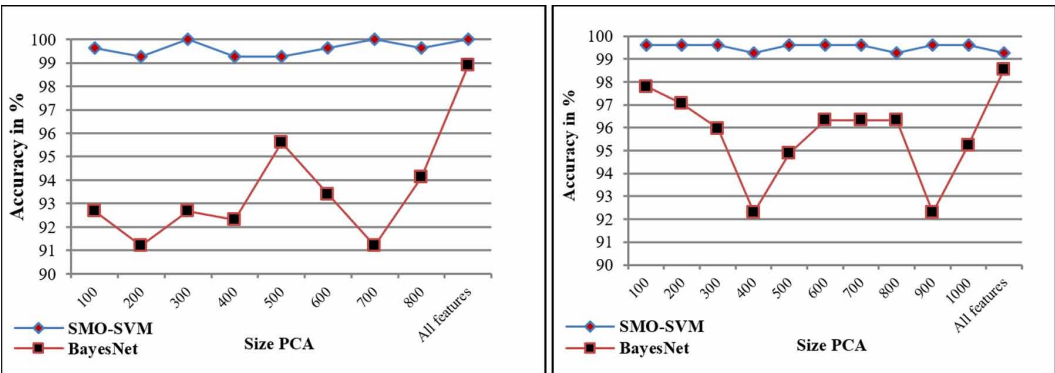
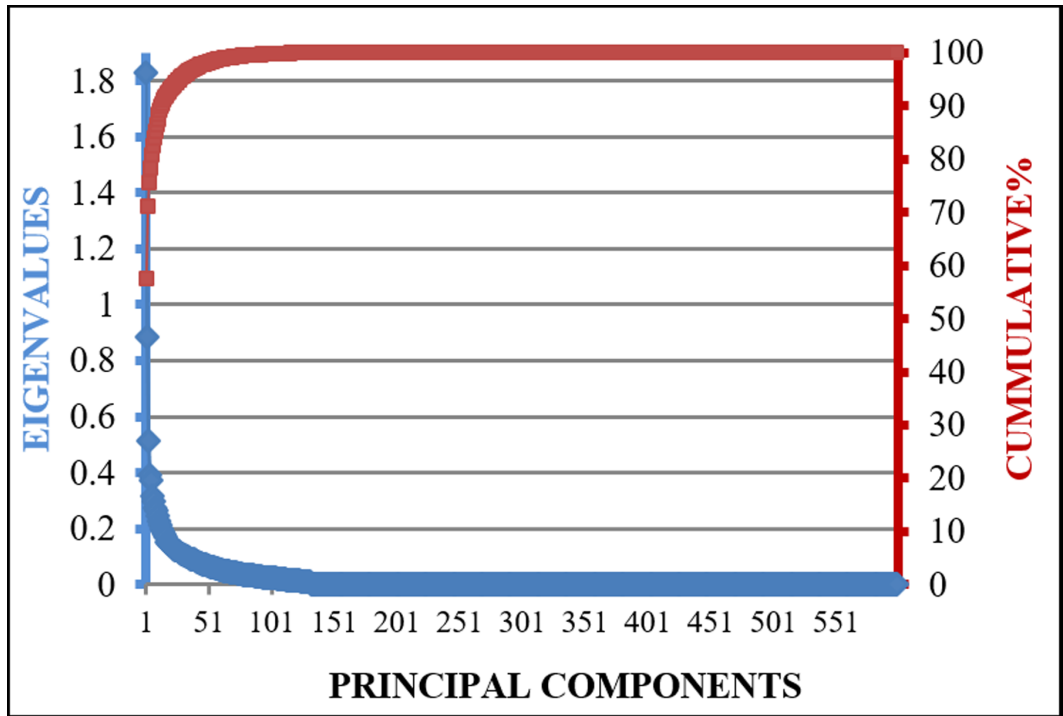


Figure 6. Eigenvalues and the cumulative variation deduced by the PCA. The case of function words



It can be clearly seen that the PCA+SMOTE improves significantly the accuracy of SVM classifier, where the best performances (100%) are obtained by Starting Bigrams. However it provides a less accuracy (about 91.5%) with NB classifier using FW.

Figure 8 shows the obtained F-measure using both techniques: SMOTE and PCA+SMOTE. We notice that the proposed method is suitable with FW, and the F-measure is increased from 95% using SMOTE to 99% using PCA+SMOTE. In addition, The SMOTE approach has produced the best performance with NB using all the features.

Figure 7. Eigenvalues and the cumulative variation deduced by the PCA. Left: The case of starting bigrams. Right: The case of starting trigrams.

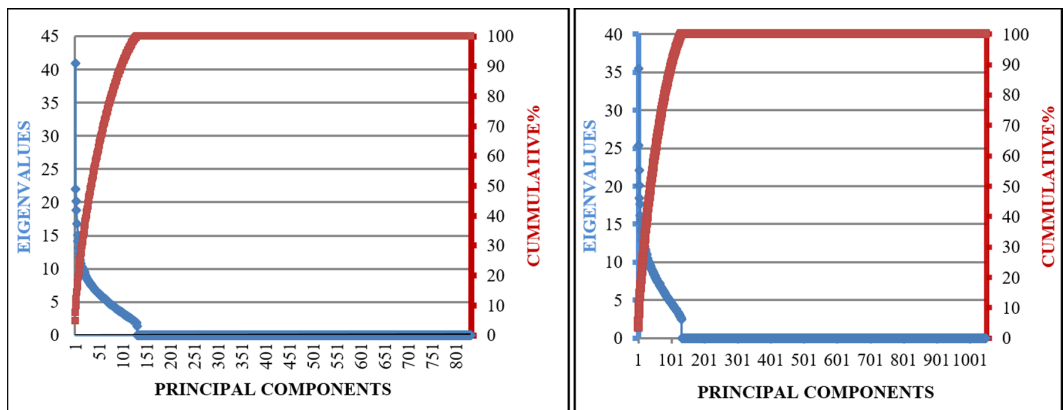


Table 4. Accuracies produced using SMOTE and SMOTE+PCA

Approach	Features					
	FW		Starting Bigrams		Starting Trigrams	
	SMO-SVM	BayesNet	SMO-SVM	BayesNet	SMO-SVM	BayesNet
SMOTE	95.97	93.77	100	98.90	99.26	98.53
SMOTE+PCA	99.66	91.57	100	95.60	99.66	96.33

The best accuracy of SVM and NB using different features across the 5-fold cross validation on unbalanced data are summarized in Figure 9. We notice from the figure that the combination of both methods (PCA and SMOTE) can lead to high performances using SVM and the latter outperforms NB classifier as expected. Finally, we conclude that even though the PCA and SMOTE can reduce the features size and improve the identification task when used alone, and the combination of both can lead to high performances.

### 5.5. Comparison With Other Methods

For comparative purposes, we considered some works conducted on AA with unbalanced Arabic datasets. Stamatatos (2008), used four methods to handle unbalanced multi-class textual datasets in AA, he segmented the training set into several samples according to the class size. The evaluation of the four methods is based on two corpora, i.e. newswire stories in English and newspaper reportage in Arabic. This work was one of the most important in Arabic AA using unbalanced dataset. Hence, we decided to compare our work with this work. Figure 10 illustrates a comparison between our results and those obtained by Stamatatos's method in terms of accuracy on our dataset (SAB-2).

The results show that the proposed method achieved the best accuracy (100%) using SVM, while the second-best accuracy was produced by Stamatatos's method (about 96.33%).

Figure 8. Comparison between SMOTE and PCA+SMOTE using different features. Left: using SMOTE. Right: using PCA+SMOTE

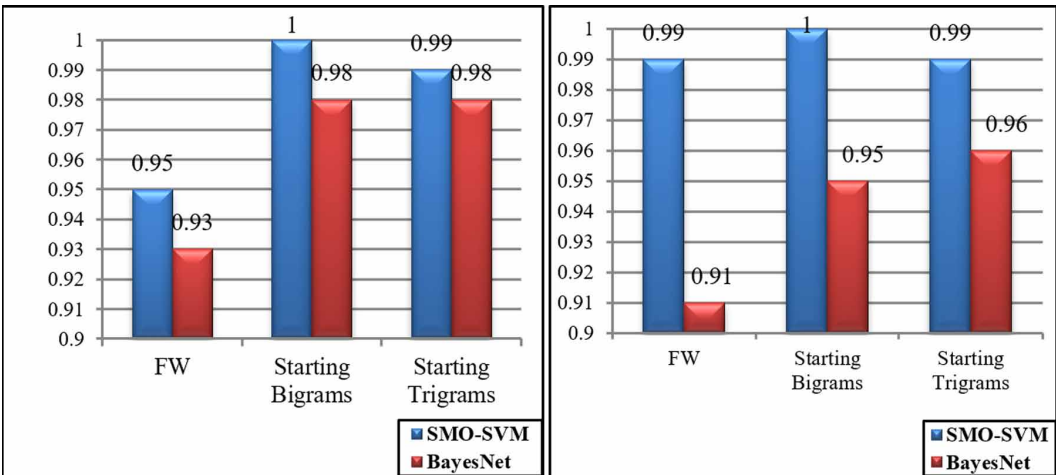
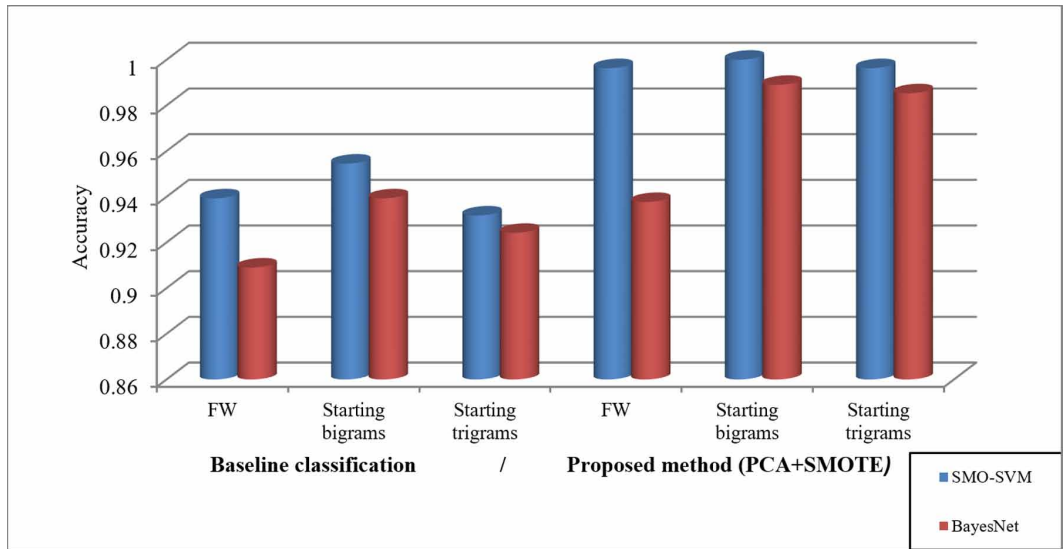


Figure 9. Comparison between the baseline and proposed method (PCA+SMOTE)

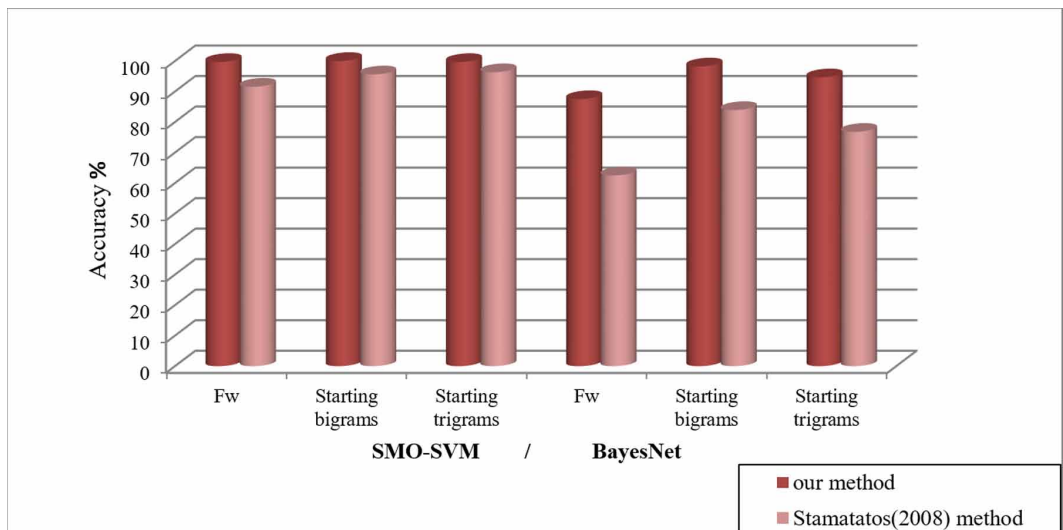


## 6. CONCLUSION

In this paper, we have proposed a new hybrid approach between PCA reduction and SOMTE in authorship attribution using unbalanced dataset. In this regard, we have conceived a new Arabic dataset (called SAB-2 corpus). In addition, we have proposed two types of features such as function words and starting n-grams. The set of features is reduced by the PCA, and the results are submitted to the SMOTE part to construct a balanced data. We have used two classifiers, namely: SVM and NB, where they are evaluated based on the cross-validation because of the limited dataset size (SAB-2).

We have conducted three sets of experiments, where the first one concerns the classification without preprocessing. The second experiment aims to study the effect of the feature set size, and

Figure 10. Comparison of our method proposed with the methods of Statamatos



the last one focuses on studying the effect of our proposed feature selection method. We can draw the following conclusions from the conducted experiments:

- Without preprocessing, SVM can reach 95.45% of accuracy using starting character tri-grams, while NB classifier produces the lowest accuracy (about 90.90%) with function words.
- PCA can select nearly a half of the global feature set when we use function words and starting bigrams, and can select one third of the feature set when we use starting trigrams.
- The combination between PCA and SMOTE considerably improves the classification accuracy, where the SVM reaches 100% of accuracy using starting bigrams.
- The hybrid approach can deal effectively with unbalanced datasets.

As perspectives, we expect to extend our approach to author verification and discrimination. Moreover, we intend to extend the experiments to larger unbalanced corpora to get more significant results.

## **ACKNOWLEDGMENT**

We would like to thank warmly the editor-in-chief and the reviewers for their valuable comments. We would like also to express our sincere acknowledgements to Dr Khennouf and Dr Abainia for their help during the elaboration of this paper.

## REFERENCES

- Abbasi, A., & Chen, H. (2005). Applying authorship analysis to extremist-group Web forum messages. *IEEE Intelligent Systems*, 20(5), 67–75. doi:10.1109/MIS.2005.81
- Al-Falahi, A., Ramdani, M., & Bellafkih, M. (2019). Arabic poetry authorship attribution using machine learning techniques. *Journal of Computational Science*, 15(7), 1012–1021. doi:10.3844/jcsp.2019.1012.1021
- Alhakbani, H. (2018). *Handling class imbalance using swarm intelligence techniques, hybrid data and algorithmic level solutions* (Doctoral dissertation). University of London, London, UK.
- AlSallal, M., Iqbal, R., Palade, V., Amin, S., & Chang, V. (2019). An integrated approach for intrinsic plagiarism detection. *Future Generation Computer Systems*, 96, 700–712. doi:10.1016/j.future.2017.11.023
- Alwajeeh, A., Al-Ayyoub, M., & Hmeidi, I. (2014). On authorship authentication of Arabic articles. In *2014 5th International Conference on Information and Communication Systems, ICICS 2014*, (pp. 1-6). doi:10.1109/IACS.2014.6841973
- Argamon, S., & Levitan, S. (2005). Measuring the usefulness of function words for authorship attribution. *Proceedings of the ACHALLC*.
- Argamon, S., Whitelaw, C., Chase, P., Hota, S. R., Garg, N., & Shlomo Levitan, L. (2007). Stylistic text classification using functional lexical features. *Journal of the American Society for Information Science and Technology*, 58(6), 802–822. doi:10.1002/asi.20553
- Baayen, R. H., van Halteren, H., Neijt, A., & Tweedie, F. (2002). An experiment in authorship attribution. In *Proceedings of 6th International Conference on the Statistical Analysis of Textual Data (JADT 2002)*, (pp. 29–37). Academic Press.
- Bozkurt, I. N., Bağlioğlu, Ö., & Uyar, E. (2007). Authorship attribution: Performance of various features and classification methods. In *Proceedings of the 22nd international symposium on Computer and information sciences*, (pp. 1–5). doi:10.1109/ISCIS.2007.4456854
- Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16(1), 321–357. doi:10.1613/jair.953
- Eder, M. (2015). Does size matter? Authorship attribution, small samples, big problem. *Digital Scholarship in the Humanities*, 30(2), 167–182. doi:10.1093/lc/fqt066
- Edwards, M. (2018). *Data quality measures for identity resolution* (Doctoral dissertation). Lancaster University.
- García, S., Luengo, J., & Herrera, F. (2016). Tutorial on practical tips of the most influential data preprocessing algorithms in data mining. *Knowledge-Based Systems*, 98, 1–29. doi:10.1016/j.knosys.2015.12.006
- Genkin, A., Lewis, D. D., & Madigan, D. (2007). Large-scale bayesian logistic regression for text categorization. *Technometrics*, 49(3), 291–304. doi:10.1198/004017007000000245
- Grieve, J. (2007). Quantitative authorship attribution: An evaluation of techniques. *Literary and Linguistic Computing*, 22(3), 251–270. doi:10.1093/lc/fqm020
- Guandong, X., Zong, Y., & Zhenglu, Y. (2013). *Applied data mining*. CRC Press.
- He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284. doi:10.1109/TKDE.2008.239
- Heckerman, D., Geiger, D., & Chickering, D. M. (1995). Learning Bayesian Networks: The Combination of Knowledge and Statistical Data. *Machine Learning*, 20(3), 197–243. doi:10.1007/BF00994016
- Hoang, G., Bouzerdoum, A., & Lam, S. (2009). Learning Pattern Classification Tasks with Imbalanced Data Sets. *Pattern Recognition*, 193-208. doi:10.5772/7544
- Holmes, D. I., Robertson, M., & Paez, R. (2001). Stephen Crane and the New-York Tribune: A case study in traditional and non-traditional authorship attribution. *Computers and the Humanities*, 35(3), 315–331. doi:10.1023/A:1017549100097

Jolliffe, I. T. (2002). *Principal Component Analysis*. Encyclopedia of Statistics in Behavioral Science. doi:10.2307/1270093

Juola, P. (2004). A\_hoc authorship attribution competition. In *Proceedings of the Joint Conference of the Association for Literary and Linguistic Computing* (pp. 175-172). Academic Press.

Juola, P. (2006). Authorship attribution. *Foundations and Trends in Information Retrieval*, 1(3), 233–334. doi:10.1561/15000000005

Keerthi, S. S., Shevade, S. K., Bhattacharyya, C., & Murthy, K. R. K. (2001). Improvements to Platt's SMO algorithm for SVM classifier design. *Neural Computation*, 13(3), 637–649. doi:10.1162/089976601300014493

Kešelj, V., Peng, F., Cercone, N., & Thomas, C. (2003). N-gram-based author profiles for authorship attribution. In *Proceedings of the Pacific Assoc. for Computational Linguistics* (pp. 255-264). Academic Press.

Kjell, B. (1994). Authorship attribution of text samples using neural networks and Bayesian classifiers. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics*. IEEE. doi:10.1109/ICSMC.1994.400086

Koppel, M., Schier, J., & Bonchek-Dokow, E. (2007). Measuring Differentiability: Unmasking Pseudonymous Authors. *Journal of Machine Learning Research*, 8, 1261–1276.

Koppel, M., & Schler, J. (2003). Exploiting Stylistic Idiosyncrasies for Authorship Attribution. In *IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis* (pp. 69-72). Academic Press.

Koppel, M., Schler, J., & Argamon, S. (2011). Authorship attribution in the wild. *Language Resources and Evaluation*, 45(1), 83–94. doi:10.1007/s10579-009-9111-2

Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. In *Progress in Artificial Intelligence*, 5, 221-232. .10.1007/s13748-016-0094-0

Li, F., Zhang, X., Zhang, X., Du, C., Xu, Y., & Tian, Y. C. (2018). Cost-sensitive and hybrid-attribute measure multi-decision tree over imbalanced data sets. *Information Sciences*, 422, 242–256. doi:10.1016/j.ins.2017.09.013

Li, W., Zhao, D., Yang, J., & Cao, L. (2013). An approach of hierarchical concept clustering on Medical Short Text corpus. In *Proceedings of the 2013 6th International Conference on Biomedical Engineering and Informatics, BMEI 2013* (pp. 509 – 518). doi:10.1109/BMEI.2013.6746995

Madigan, D., Genkin, A., Lewis, D., Argamon, S., Frakin, D., & Ye, L. (2005). Author identification on the large scale. *Proceedings of CSNA-05*.

Mohd, F., Jalil, M. A., Noora, N. M. M., Ismail, S., Yahya, W. F. F., & Mohamad, M. (2019, December). Improving Accuracy of Imbalanced Clinical Data Classification Using Synthetic Minority Over-Sampling Technique. In *International Conference on Computing* (pp. 99-110). Springer. doi:10.1007/978-3-030-36365-9\_8

Mosteller, F., & Wallace, D. L. (1984). *The Federalist Papers As a Case Study*. .10.1007/978-1-4612-5256-6\_1

Ouamour, S., & Sayoud, H. (2018). *A Comparative Survey of Authorship Attribution on Short Arabic Texts*. Lecture Notes in Computer Science. Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics. doi:10.1007/978-3-319-99579-3\_50

Pan, T., Zhao, J., Wu, W., & Yang, J. (2020). Learning imbalanced datasets based on SMOTE and Gaussian distribution. *Information Sciences*, 512, 1214–1233. doi:10.1016/j.ins.2019.10.048

Peng, F., Shuurmans, D., Keselj, V., & Wang, S. (2003). Language independent authorship attribution using character level language models. In *Proceedings of the 10th Conference of the European chapter of the Association for Computational Linguistics* (pp. 267-274). doi:10.3115/1067807.1067843

Rocha, A., Scheirer, W. J., Forstall, C. W., Cavalcante, T., Theophilo, A., Shen, B., & Stamatatos, E. (2017). Authorship Attribution for Social Media Forensics. *IEEE Transactions on Information Forensics and Security*, 12(1), 5–33. doi:10.1109/TIFS.2016.2603960

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5), 513–523. doi:10.1016/0306-4573(88)90021-0



- Sari, Y., Vlachos, A., & Stevenson, M. (2017). Continuous n-gram representations for authorship attribution. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, 2(2), 267–273. doi:10.18653/v1/E17-2043
- Shaker, K., & Corne, D. (2010). Authorship attribution in Arabic using a hybrid of evolutionary search and linear discriminant analysis. In *2010 UK Workshop on Computational Intelligence*. IEEE. doi:10.1109/UKCI.2010.5625580
- Soltanzadeh, P., & Hashemzadeh, M. (2021). RCSMOTE: Range-Controlled synthetic minority over-sampling technique for handling the class imbalance problem. *Information Sciences*, 542, 92–111. doi:10.1016/j.ins.2020.07.014
- Stamatatos, E. (2006a). Authorship attribution based on feature set subsampling ensembles. *International Journal of Artificial Intelligence Tools*, 15(5), 823–838. doi:10.1142/S0218213006002965
- Stamatatos, E. (2006b). Ensemble-based author identification using character n-grams. In *Proceedings of the 3rd International Workshop on text-based Information Retrieval* (pp. 41–46). Academic Press.
- Stamatatos, E. (2008). Author Identification: Using Text Sampling to Handle the Class Imbalance Problem. *Published in Information Processing and Management*, 44(2), 790–799. doi:10.1016/j.ipm.2007.05.012
- Stamatatos, E., Fakotakis, N., & Kokkinakis, G. (2001). Computer-based authorship attribution without lexical measures. *Language Resources and Evaluation*, 193–214.
- Swathi, C., Karunakar, K., Archana, G., & Raghunadha Reddy, T. (2018). A new term weight measure for gender prediction in author profiling. *Advances in Intelligent Systems and Computing*, 695, 11–18. doi:10.1007/978-981-10-7566-7\_2
- Weiss, S. M., & Kulikowski, C. A. (1991). Computer systems that learn: classification and prediction methods from statistics Nets, Machine Learning, and Expert Systems. In *Neural Networks. Machine Learning, and Expert Systems*.
- Yule, G. U. (1944). The statistical study of literary vocabulary. *Modern Language Review*, 39(3), 291. Advance online publication. doi:10.2307/3717870
- Zhang, D., & Lee, W. S. (2006). Extracting key-substring-group features for text classification. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 474–483). doi:10.1145/1150402.1150455
- Zhang, J. (2019). Machine Learning With Feature Selection Using Principal Component Analysis for Malware Detection: A Case Study. In *Sophos technical papers* (pp. 1–5). Sophos.
- Zhao, Y., & Zobel, J. (2005). Effective and scalable authorship attribution using function words. In *AIRS'05 Proceedings of the Second Asia conference on Asia Information Retrieval Technology* (pp. 174–189). Springer. doi:10.1007/11562382\_14