Available online at www.jlls.org

JOURNAL OF LANGUAGE AND LINGUISTIC STUDIES

ISSN: 1305-578X

Journal of Language and Linguistic Studies, *17*(1), 542-557; 2021

Authorship attribution of Morsi Gameel Aziz's lyrics: A clustering-based stylometry approach

Abdulfattah Omar [1]

Prince Sattam Bin Abdulaziz University, Alkharj, Saudi Arabia
Port Said University, Port Said, Egypt

APA Citation:

Omar, A. (2021). Authorship attribution of Morsi Gameel Aziz's lyrics: A clustering-based stylometry approach. *Journal of Language and Linguistic Studies*, *17*(1), 542-557. Doi: 10.52462/jlls.36

Submission Date: 12/01/2021

Acceptance Date: 16/03/2021

Abstract

Numerous studies have addressed the issue of the authorship of Morsi Gameel Aziz's lyrics. These studies have traditionally been based on chronological criteria for determining the real authors of disputed lyrics. To date, there is no agreement on the real authors of these disputed lyrics. This can mainly be attributed to both selectivity and the lack of empirical evidence in such studies, raising questions about the reliability of such approaches. With the advent of machine learning systems and data mining techniques, it is now possible to process thousands of texts using replicable methods. Thus, this study seeks to address the issue of the authorship of Morsi Gameel Aziz's lyrics making use of these advances by applying a clustering-based stylometry approach. The hypothesis is that lyrics grouped or clustered together are more likely to be written by the same poet. A corpus of 1,089 lyrics was built, including all known lyrics attributed to Aziz and the lyrics of the poets thought to be the real authors of the disputed lyrics. The lyrics were clustered using the Gibbs sampling Dirichlet multinomial mixture (GSDMM) technique and were assigned to four main classes, with the 12 disputed lyrics clustered within Aziz's class. Based on this, it is clear that the GSDMM model is effective and reliable in clustering short documents in Arabic. The results of the study show that machine learning systems and stylometric authorship techniques can be used in resolving many authorship questions that remain controversial and unanswered in Arabic literature.

Keywords: authorship attribution; clustering; Gibbs sampling Dirichlet multinomial mixture; letter pairs; lyrics; Morsi Gameel Aziz; stylometry

1. Introduction

Literary thefts are as old in the history of human thought and literature as mankind. The first known incidents can be traced back to ancient Greeks and Romans. Aristotle referred to literary thefts when he mentioned that poets used ancient expressive images, quoting from their ancient counterparts. Horace, the Latin poet, admitted that he imitated many of his predecessors. However, at this time, imitations were considered natural and there were no ethical or legal issues concerning such practices.

[1] Corresponding author.
E-mail address: a.abdelfattah@psau.edu.sa

It was not until the 17th century that people became more concerned with originality as they began to place a high value on original works, rather than imitations. During this time, the English poet and playwright Ben Jonson coined the word ‘plagiary’ to denote literary theft, paving the way for considering plagiarism as a crime in the 18th century (Pask, 2002). Indeed, literary theft is now considered one of the most important ongoing problems in world literature. Authorship questions remain among the most controversial topics in literary circles.

In Arabic literature, many poets, including Zuhair bin Abi Salma, Tarfa ibn al-Abd, and Abu al-Tayyib al-Mutanabbi, have long been accused of plagiarism. Even worse, various questions concerning authorship have been raised about pre-Islamic poetry which is considered the most celebrated form of Arabic literature. In his book *On the Pre-Islamic Poetry*, Dr. Taha Hussein (1927), one of the pioneers of modern Arabic literature, argued that pre-Islamic poetry was transformed and rewritten after Islam, and it was then attributed to pre-Islamic poets. Hussein's book provoked controversy, which made it one of the most famous books of the 20th century. Hussein’s arguments gave a strong shock to his contemporaries, one that continues to resonate today.

In the second half of the 20th century, the issues of plagiarism and authorship attribution became a preoccupation among many scholars and critics. Mustafa Sadiq al-Rafi'i accused Abbas Al-Aqqad of literary theft, referring to it as litigation and fraud. Abbas Mahmoud Al-Akkad was a writer, poet, and literary critic. He was also a philosopher, politician, journalist, and historian. Al-Aqqad is remembered as one of the best Arab writers of all time. The issue received much attention at the time.

The issue of literary theft did not stop with poetry and novels, but rather it extended to cinematic and televised dramas, melodies, songs, and lyrics. Over the years, different critics have accused Morsi Gameel Aziz of stealing his lyrics, and he was even once described as the pirate of lyrics. Aziz (1921–1980) was an Egyptian poet and songwriter; his father was one of the major fruit merchants and enabled Aziz to satisfy his literary and artistic interests when he was young. Aziz was influenced by the vendors' calls for fruit and folk songs. His tendency toward poetry appeared early, specifically at the age of 12. When he was only 19, he started a long vocation in poetry, writing lyrics to be sung by top singers accompanying the works of great composers. He wrote more than one thousand lyrics. For many critics, he represented a milestone, not only in his creative career but also in Arab cultural life in general.

In 1961, Galeel Albendary, a critic who was famous for his harsh criticism of poets and artists, accused Aziz of stealing 12 lyrics including his famous lyrics *Ya ma Alqamar 'ealbab Yama Anadyloh (O Ma, Moon-Like Sweetheart Is by the Door!)*, *Hubak Nar* (*Your Love is Glowing in My Heart. Fire of Passion!)*, *Btlwmwny Lyh law shwftom 'eynyh* (*Blame Me Not; Why on Earth Would You?, Seen, Not Told!)*, and *Ana shwft Jamal* (*Ma, I Beheld the Beauty (of Nasser)!*). Albendary even described Aziz as the lyrics pirate. These accusations led to the formation of a critical jury of leading academics and critics to adjudicate the theft cases in a way that was unprecedented at this time. After the investigation and examination of the evidence and documents, the jury issued its verdict acquitting Morsi Gameel Aziz of the charge of plagiarism in at least 10 poems and accusing him of quoting 2 lyrics. However, the decision of the jury did not satisfy Aziz, his opponents, or the public; therefore, the issue was not resolved. Since then, the issue has been investigated in different studies, which have generally used criteria for attributing works on chronological and epistemological bases to determine the real authors of disputed lyrics. So far, there has been no agreement on the real authors of these disputed lyrics. This can mainly be attributed to both selectivity and lack of empirical evidence in such studies, raising questions about the reliability of used approaches. With the advent of computer technology and linguistic stylometry methods, it is now possible to process thousands of texts using replicable methods. Thus, this study asks whether the authorship question of Morsi Gameel Aziz’s lyrics can be resolved using an authorship attribution technique based on indexing letter-pair patterns.

The rest of the paper is organized as follows. Part 2 surveys the authorship attribution and stylometry literature. Part 3 describes the methodological framework of the study. Part 4 reports the quantitative and statistical results of the study. Part 5 summarizes the key findings and offers avenues for further research.

2. Literature review

Authorship attribution is broadly defined as the process of determining the real author of a work that is unknown or disputed (Bagavandas & Manimannan, 2008; Iqbal, Debbabi & Fung, 2020; Juola, 2008; Love, 2002). Put simply, it is a process of assigning authorship. Authorship attribution has been traditionally based on chronological and epistemological methods. With the development of stylistic analysis, it has become possible to assign disputed texts or those written by unknown authors to their real authors through objective methods based on internal evidence, as represented in the identification of patterns formed in the process of the linguistic encoding of information. The underlying principle of stylistic approaches is that everyone has a unique style by which he/she can be identified through identifying the distinctive linguistic and stylistic choices that capture authors' writing style (Burrows, 2002; Coyotl-Morales, Villaseñor-Pineda, Montes-y-Gómez & Rosso, 2006; Hoover, 2004). Stylistic analysis, which focused on uses of language that are unique at the individual level, has become an intrinsic aspect of author attribution. Each author has a style that is as identifiable as a signature, and analysis reveals linguistic features and patterns that indicate the true author.

With the advent of computer technology, authorship attribution has come to be based mainly on quantitative and statistical methods used to identify the salient features of a text to draw accurate and reliable conclusions about its composition (Craig & Kinney, 2009; Holmes, 1995a, 1995b; Hoover, Culpeper & O'Halloran, 2014). The underlying principle has been that through the use of computers, it is possible to obtain more accurate attribution results through quantitative rigorous and intense investigations of the linguistic and stylistic patterns of authors.

The rapid advancements in quantitative investigations aimed at resolving controversial authorship problems have helped to develop quantitative authorship analysis as a distinct discipline of knowledge. This has come to be known as non-traditional authorship attribution, stylometric authorship attribution, or simply authorship attribution (Bagavandas & Manimannan, 2008; Dauber Jr, 2020; Varela, Albonico, Justino & de Assis, 2020), that is defined as "the attempt to reveal the authors behind texts based on a quantitative analysis of their style" (Stamatatos et al., 2016).

The relation between quantitative aspects and literary phenomena is very old (Omar, 2021). Numerous studies have attempted to explain the stylistic and linguistic properties of authors through quantitative methods, which have become more sophisticated with the availability of computational – rather than non-computational – methods (El Bakly, Darwish & Hefny; Omar, Elghayesh & Kassem, 2019). Many studies agree that the development of computational methods has enhanced the efficiency and accuracy of stylometric studies, since computer systems have capacities for analyzing large quantities of data (Ison, 2020; Omar & Hamouda, 2020; Zhao, Li, Qi & Da Xu, 2020). Thus, "[C]omputational stylistics provides a valuable methodology in helping to answer the question, 'Who wrote it?'" (Crabb, Antonia & Craig, 2014, p. 177).

The chief merit of quantitative analysis is that it is objective (Abu Rabiah, 2020). Besides, it can deal with large datasets effectively and rapidly. Balossi adds that quantitative methods in literary studies are valid, reliable, and effective (Balossi, 2014). Similarly, Jockers (2014) argues that with many traditional fundamental questions concerning authorship attribution and the increasing feasibility of finding answers based on the availability of computer-based technology, it is becoming imperative to go beyond traditional techniques for dealing with such issues. Craig and Kinney (2009) even

contend that the investigation of authorship of disputed or controversial texts relies more on statistical analysis than on literary investigation.

Despite the effectiveness of stylometry-based approaches in authorship studies, the issue of the reliability of such methods is still raised by some commentators. This can be attributed to the persistent wide gap between computational methods and literary studies (Schreibman, Siemens & Unsworth, 2016). Stylometry is often met with objections from many critics. They argue that the computational approach of stylometry never gives results that can be universally accepted as definitive. Rudman (Rudman, 1997, 2012) argues that the application of computational approaches in authorship attribution is neither reliable nor well-understood. He points out that stylometric authorship attribution is often blamed for the fact that it cannot be appropriately applied to other genres or languages. He adds that the question of who wrote a given text remains exactly as it was before any stylometric approach is undertaken. Holmes (1998) argues that there are two main problems with stylometry that hinder its acceptance within humanities scholarship. First, there is no consensus concerning the correct methodology or technique. Second, "no stylometrist has managed to establish a methodology which is better able to capture the style of a text than that based on lexical items." In the face of such opposing views, this study is aligned with the results of the many studies indicating that computational stylometric studies have had reasonable success in identifying the linguistic and stylistic characteristics of many authors, and even in confirming the results of conventional criticism. Related literature indicates that an impressive body of computer-based work has been done on authorship studies over the last five decades using stylometric approaches (Savoy, 2012). It is also clear that machine learning systems and data mining techniques have been employed usefully in resolving many authorship questions that had long remained unanswered in literary studies (Holmes, 1995a; Zhu, Lei & Craig, 2020).

In stylometric authorship studies, researchers have generally used various statistical multivariate analysis techniques that range from frequency distribution (i.e. listing frequently used words) to discriminant analysis examining linguistic and stylistic features within texts that can be detectors of potential authors (Daelemans, 2013; Eder, Piasecki & Walkowiak, 2017; Gómez-Adorno, Posadas-Duran, Ríos-Toledo, Sidorov & Sierra, 2018; Hoover, 2003; Lagutina, Boychuk, Vorontsova & Paramonov, 2019). These statistical techniques have been used with different linguistic variables, including morphological, lexical, and syntactic variables (Burrows, 2007; Strome, 2013).

Stylometry-based approaches have been extensively used in studies to resolve many authorship issues in various world pieces of literature (Savoy, 2012; Zhu et al., 2020). The questions surrounding the authorship of the English dramatist and poet William Shakespeare remain the most highly investigated in stylometric authorship studies (Craig & Greatley-Hirsch, 2017; Craig & Kinney, 2009; Mathews & Merriam, 1993; Smith, 1992). In these studies, new methods have been developed to answer long-standing questions and challenge the fact that traditional or non-computational literary accounts have failed to give answers.

Despite the effectiveness of stylometric approaches, machine learning systems, and data mining techniques and their potential in resolving different authorship questions, these approaches and systems have not been applied properly in authorship studies addressing Arabic literature. Authorship attribution of Arabic literary texts has so far largely been based on traditional and non-computational methods. Only in recent years have some researchers become interested in applying learning systems and data mining techniques to the study of authorship in Arabic.

In an analysis of the authorship attribution of flash fiction in Arabic, Omar et al. (2019) pointed out that authorship questions have dramatically increased over recent years due to the unprecedented development of online social platforms and the difficulty of handling authorship issues through

traditional methods. They suggested a morphological-based stylometric technique for the identification of real authors. Similarly, vector space clustering methods have successfully been used to identify the authorship of Arabic poetry (Al-Falahi, Ramdani & Bellafkih, 2017, 2019). Although these studies show the potential of computational stylometry, data mining, and machine learning systems in authorship inquiry, there is still a lack of use of computational tools in Arabic authorship studies. This study seeks to address this gap in the literature through addressing the authorship issue of Morsi Gameel Aziz's lyrics using a stylometry-based approach.

3. Methodology

To assign the lyrics to their true poets, the study applied document clustering based on the Gibbs Sampling Dirichlet Multinomial Mixture (GSDMM) technique. Document clustering was chosen as the methodological framework as it is appropriate for organizing any unorganized set of documents (Aggarwal & Reddy, 2018; Wu, Xiong & Shekhar, 2013). It has been widely used in different authorship applications to group semantically related documents, and it has been found to give reliable results. Indeed, document clustering methods and techniques are used for automatic clustering and classification of text documents in a variety of areas, including adaptive information filtering, information distillation, and text search (Srivastava & Sahami, 2009). In stylometric authorship attribution studies, the task of identifying the real or correct authors of disputed texts has always entailed document clustering. According to Zheng and Zheng (2020), "the general approach to authorship attribution is to extract a set of style characteristics from the text and use these characteristics as features to train a classifier" (Zheng & Zheng, 2020, p. 321). In other words, authorship attribution is primarily a clustering task.

The underlying principle of document clustering theory is that machine learning systems can be used to structure unorganized datasets and group them into distinct classes, known as clusters, by computing similarity using multivariate analysis methods, most notably cluster analysis. In mathematical terms, clustering techniques are used to discover related sub-spaces in a multi-dimensionally distributed dataset (Wu et al., 2013). In other words, vectors can be distributed in an *n*-dimensional space in a non-random way to be assigned into distinct clusters or groups.

The distinction between random and non-random generation of vectors in terms of distribution in a multi-dimensional space and the potential for clustering is illustrated in Figures 1 and 2. Figure 1 presents an example of a plot of 100 three-dimensional randomly generated vectors. It demonstrates that the random generation of data results in non-uniform distribution, making it difficult to identify relations among the vectors beyond some weak associations.

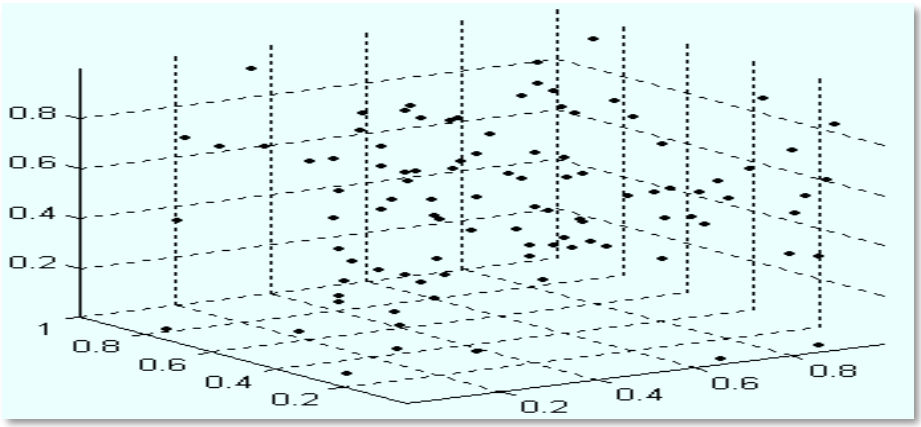

Figure 1. Plot showing 100 three-dimensional randomly generated vectors

In contrast, if vectors are organized into a non-random dimensional space, as shown in Figure 2, it is possible to define relations among the vectors.

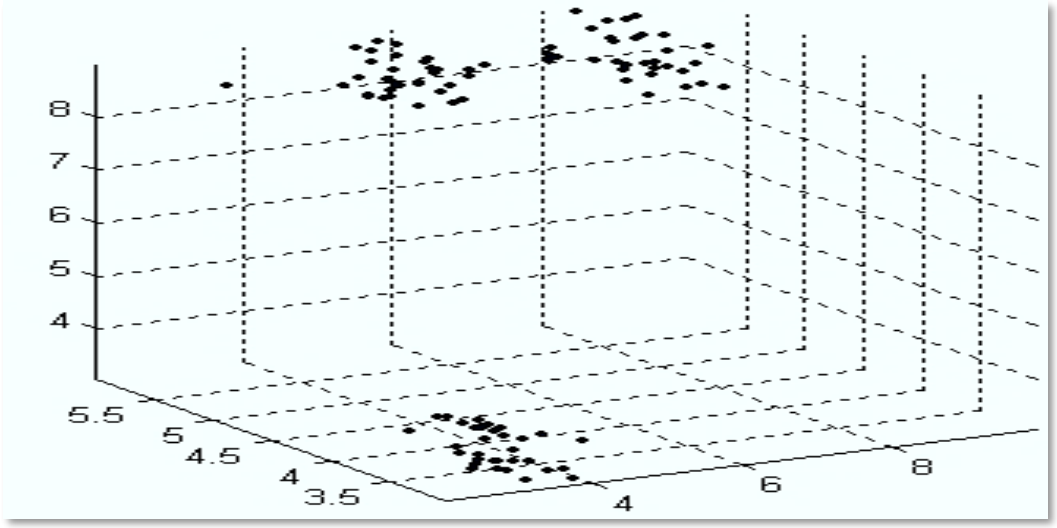

Figure 2. Plot showing 100 three-dimensional non-randomly generated vectors

Unlike the random distribution of the dataset shown in Figure 2, the distribution of points in this example is non-random, i.e. the vectors are assigned to three clearly defined groups. However, because datasets are usually very large, it is difficult to detect regularities within them. Document clustering methodology offers various techniques for identifying data clusters and presenting them graphically in a way that is intuitively and observable. The reasons for selecting document clustering concern two aspects: objectivity and appropriateness.

Document clustering is generally used in the absence of prior hypotheses about relations in the data, thus serving objectivity, the main concern of this research. Document clustering is a means of dealing effectively with data that are large in terms of the number of objects being studied for dimensionality. In this case, it makes the methodology useful in generating a classification of lyrics, which can then be applied to make hypotheses about the true authors of the disputed texts.

Document clustering is associated with multivariate analysis (MVA). As its name indicates, MVA entails the analysis of data with multiple dimensions or variables. It is used to describe interrelationships in data sets that have numerous variables. It employs several techniques, ranging from correlation analysis, factor analysis, discriminant analysis, and correspondence analysis to principal components analysis (PCA) and cluster analysis (Moisl, 2015; Srivastava & Sahami, 2009).

A variety of classifications have been suggested about MVA. However, statisticians view MVA as comprising a combination of two complementary methods: those that seek to discover structure from the evidence provided by the data matrix alone and those that assume a given structure. In technical terms, MVA techniques can be generally classified under two main headings: exploratory (E)MVA and confirmatory MVA (Moisl, 2015). Generally speaking, EMVA is used to form hypotheses about the data, while confirmatory MVA is used to determine whether there are relations between some number of selected independent variables and one or more dependent variables. In other words, EMVA techniques are used to generate hypotheses about the data, while confirmatory techniques are used for hypothesis testing purposes (Timm, 2007). Hence, EMVA methods can be considered descriptive in nature, in that they are simply concerned with describing what is going on in the data, whereas confirmatory methods are inferential in that they make judgments and reach conclusions that extend beyond the immediate data alone (Denis, 2020; Everitt & Hothorn, 2011).

By way of demonstration, assume a dataset comprising a population of 30,000 people from 40 world cities representing different professional categories. With this dataset, we can undertake several MVA tasks, some of which are addressed here. If we simply want to obtain a summary or an overview of the data, we can employ PCA or factor analysis as an exploratory task, which is thus classified as EMVA. If we want to distribute the data sets into discrete groups or clusters based on their similarity/dissimilarity, we use cluster analysis, which is again EMVA. If we seek to find useful patterns in the data, often referred to as pattern recognition or data mining, it is again an EMVA task. However, if we want to prove/disprove a preconceived hypothesis about the data sets (let us say the number of doctors and poor cities or the number of architects and ancient historic cities), the analysis is confirmatory. This study is concerned with the former, EMVA, because it seeks to discover the underlying structure of the data without making a priori assumptions or theorizing about the authorship of the lyrics.

EMVA comprises a bundle of many mathematical and statistical techniques used to understand the interrelationships in data. It aims to find similarities and/or dissimilarities within the data to formulate hypotheses about the domain of interest (Moisl, 2015). EMVA begins with a problem, uses data, provides analysis, generates hypotheses about the data from the analysis, which can be called a "model," and finally draws conclusions about the data, establishing whether the hypotheses are valid. With its descriptive nature, EMVA is used for a variety of applications that range from reducing data sets with the purpose of best describing them, discovering an underlying structure within data, identifying important variables, and clustering data into classes or groups to determining optimal factor settings (Everitt, 2009). Clustering can be done through different machine learning systems. The traditional method is vector space clustering (VSC), commonly termed "a bag of words approach" as it treats text as a string of words, without considering the context or word order. The approach is widely used due to its conceptual simplicity and ease of determining semantic similarity within documents (Zhiguo, Luo, Chen, Wang & Lei, 2011). One problem with VSC, however, is that it cannot deal with short documents effectively due to sparsity (Amensisa, Patil & Agrawal, 2018; Moisl & Maguire, 2008; Omar & Aldawsari, 2019). Given the nature of the lyrics in this study, the GSDMM technique developed by Yin and Wang (Yin & Wang, 2014) was selected. The rationale is that GSDMM has been shown to be reliable in clustering short texts (Linwei Li, Guo, He, Jing & Wang, 2019; Mazarura & Waal, 2016). As Yang, Haung, and Cai (2019, p. 92040) note:

GSDMM is a probabilistic generative model for short text corpus. It relieves the sparse issue of short texts in clustering tasks with the assumption that each short text is generated by a single latent topic. The topics of short texts are inferred through Collapsed Gibbs sampling methods and used as cluster labels.

The GSDMM technique provides complete information about the documents under consideration and the degree of overlap between each pair of documents, which is useful in determining the similarity between them. It also addresses the issue of variation of document length when performing clustering. Furthermore, GSDMM potentially exhibits sound performance in incremental clustering (Yin & Wang, 2014).

4. Data Processing & Analysis

In clustering-based stylometric authorship applications, it is necessary to make the appropriate selection of a data representation scheme, as well as a clustering technique. It is generally agreed that generating prepared and adjusted datasets is crucial to obtain reliable clustering results. Variables must be selected with care because clustering algorithms have no mechanisms for differentiating between relevant and irrelevant variables. It is the job of the researcher to select all and only the variables that

will help in generating meaningful clusters. It is also important to choose the right clustering algorithm to generate reliable clustering structures. As there are no universal rules for selecting the best clustering algorithm, it is the task of the researcher to identify which clustering technique works best with the dataset. The procedures in data processing and analysis include (1) data collection and preparation, (2) data representation, (3) feature selection, (4) data clustering, and finally (5) profile alignment. These procedures are shown in Figure 3.

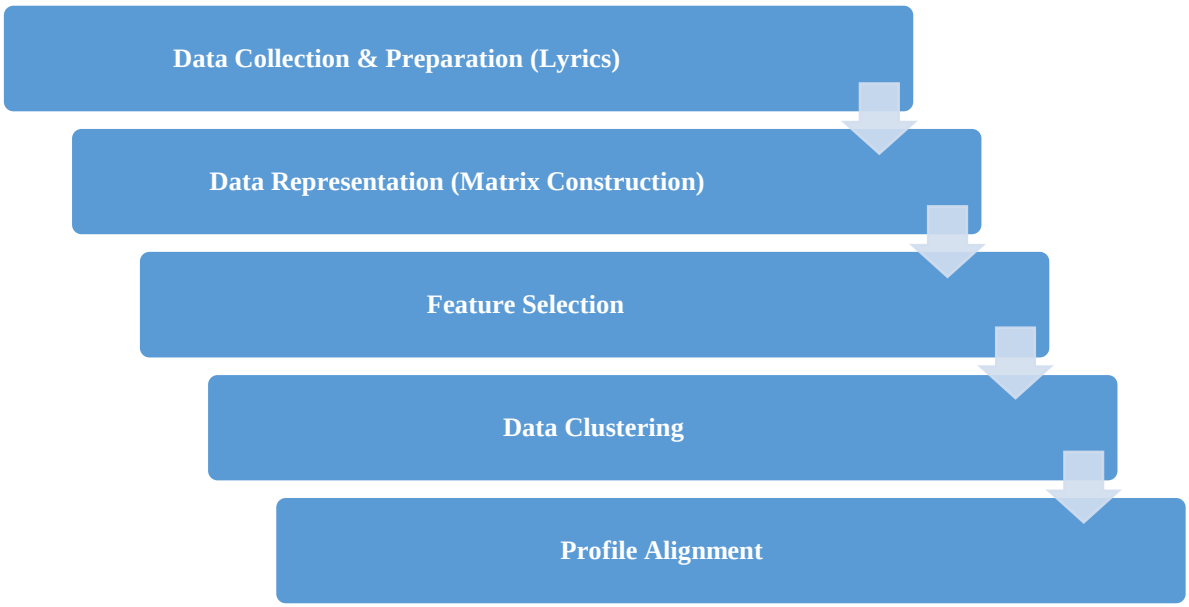

Figure 3. Data processing and analysis procedures

4.1. Data collection & preparation

Traditionally, building a corpus for text clustering applications is based on the assumption that the corpus is large and representative of the research domain. Thus, a relevant issue in this context is what size the corpus should be to support objective and reliable generalizations about the authorship of Morsi Gameel Aziz's lyrics. As an initial step, a corpus of 1,089 lyrics, including all Aziz's known lyrics and the lyrics of the poets thought to be the real authors of the disputed lyrics, was built. All Aziz's lyrics (around 1,000) have clear authorship. Only 12 lyrics have been claimed to be stolen or plagiarized. All the lyrics were downloaded from the online Arab digital archive Aghany Lyrics (available on http://aghanilyrics.com). All extraneous materials and non-alphabetic characters were then removed. These are standard preprocessing procedures in document clustering applications. Although stemming is one such procedure, it was not executed in this study as the unique stylistic features of affixation in Arabic make it possible to enhance detection and attribution of authorship using morphological information (Omar et al., 2019; Omar & Hamouda, 2020).

4.2. Data representation

An essential step in data processing is to make the corpus amenable to mathematical and statistical analysis. This is referred to as data representation, and it is required in all document clustering applications in which it is not possible for raw texts to be computationally processed. Texts must be in a form amenable to a consistent and structured analysis. In other words, a mathematical form needs to be generated so that documents are represented in a manner the computer program can handle.

To accomplish this, texts were segmented into a list of tokens representing the texts, and a matrix is constructed solely comprising the lexical types of the documents/lyrics, using what is traditionally

known as the "bag of words" technique. The lists or strings of words were then converted into letter pairs using a sequence of symbols termed a tuple: a 1-tuple is a singleton, a 2-tuple is a pair, a 3-tuple is a triple, a 4-tuple is a quadruple, and so on (Moisl, 2008). In all, 13,347 letter pairs were found. The rationale for this approach is that in authorship attribution, content words can be misleading as it is common for poets to address the same topic. Letter pairs, however, can be discriminators as each poet will have his/her own way of structuring the words used. The arrangement of words thus reflects the style and tone of each individual poet. The ways in which words are structured in poetry are closely related to the poetic craft and art, especially in Arabic, which is known for its rich and diverse morphological system.

Because the matrix is large, the job of extracting data from it becomes just like finding and catching fish in the ocean, a metaphor given by Pyle in his description of the process of data preparation (Pyle, 1999). One problem of having a matrix with such a large number of entries is that the information in it appears diffuse and redundant, making it impossible to identify the important and distinctive words that should be extracted from the corpus. One way of getting the fish out of the water is to clear the water and remove the murk so that the fish can clearly be seen (Pyle, 1999). This is referred to as feature selection.

4.3. Feature selection

This stage represents the extraction of variables that are thought to be important to ensure optimal clustering performance. It is broadly agreed that the selection of important terms is a critical task for any clustering application (Watt, Borhani & Katsaggelos, 2020). The success of any clustering depends in the first place on identifying the most important variables or features within a corpus collection. Feature selection is important for two reasons: First, not all entries are important for clustering, and second, the retention of unimportant variables can lead to unreliable clustering structures (Burges, 2010).

In clustering applications, the process of identifying and selecting the most important terms is long and complicated. In a data matrix with a large number of vectors, the data appear unclear, clouded, and redundant. Accordingly, it would be impossible to generate consistent clustering structures (Gan, Ma & Wu, 2020). The problem becomes more complicated if we consider that in clustering applications, an analyst usually works with a huge number of texts, making it difficult initially to identify what the most important terms might be. It is therefore proposed that the data matrix dimensionality be reduced as much as possible. Data extraction should be confined to the most important variables only (Hennig, Meila, Murtagh & Rocci, 2015).

Various techniques have been developed for reducing the high dimensionality of data. These have been designed to retain only the distinctive variables and remove unimportant ones. In clustering-based stylometric authorship applications, PCA and term frequency-inverse document frequency (TF–IDF) are among the most widely used techniques.

PCA is a basic tool used to reduce dimensionality in large datasets containing many interrelated variables by describing the rows and columns in a multivariate data matrix in a lower-dimensional form. It identifies the vectors that are mostly informative (Jackson, 2005), structuring the dataset without overly sacrificing useful variation, i.e., retaining data quality. Not only does PCA highlight the most distinctive variables, but it can also reveal hidden structures (Jolliffe, 2006). As such, it is commonly used in studies employing clustering techniques as it makes it possible to focus on the vectors of greatest interest.

Another common method for selecting features in clustering applications is TF–IDF, which identifies the importance of terms or words in a text (a set of documents, corpus, etc.), employing the

notion of specificity (Roelleke, 2013). It is essentially a weighting process, with the term frequency (TF) increasing in value the more the query term appears in a document, set against the number of documents in which the query term appears (IDF). It thus aids in reducing data dimensionality by identifying the most discriminant terms, hence making it possible to eliminate redundant variables (Weiss, Indurkhya, Zhang & Damerau, 2010). The results of TF–IDF can then be used for clustering analysis.

Despite the effectiveness of both PCA and TF–IDF demonstrated in different clustering applications in terms of providing reliable solutions that can reduce the high dimensionality of data in long texts, their accuracy for short texts is sub-optimal. Thus, the study employed GSDMM (see Figure 4) based on the rationale that it was an appropriate technique for data analysis as it would not only identify important information in the individual texts but also the extent of overlap (and thus similarity) between each pair of documents. Furthermore, using this technique remedies the issue of varying document lengths when undertaking to the cluster. Finally, as noted by Yin and Wang, GSDMM has good potential for use in incremental clustering (Yin & Wang, 2014).

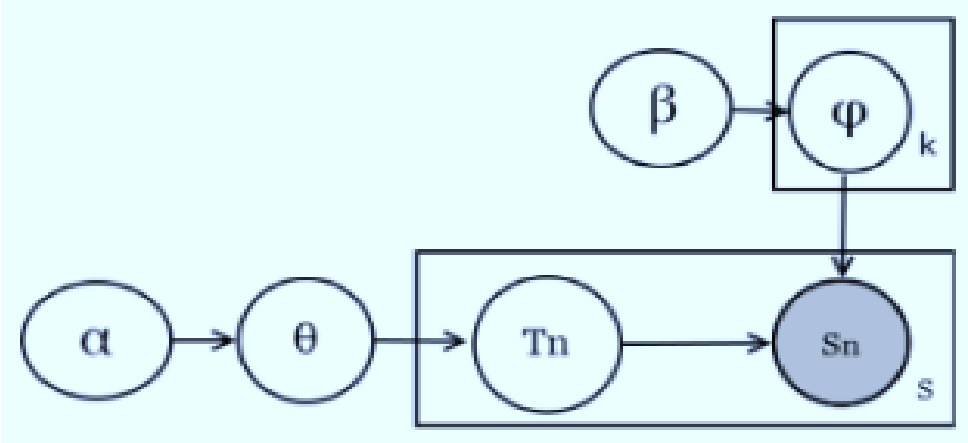

Figure 4. An illustration of the GSDMM model (Yin & Wang, 2014, p. 235)

Using the GSDMM model has the effect of reducing dimensionality and retaining only the most distinctive variables. The documents (in this case the lyrics) will be thus clustered in accordance with the letter pair patterns, and the frequency and distribution of the letter pairs. To put it simply, if letter pairs in one of the disputed lyrics present relative frequencies that are close to the letter pair distribution of a particular author, it can be concluded that the disputed lyric was written by one author.

4.4. Data Clustering

For data clustering purposes, agglomerative clustering was selected. This is based on the assumption that the greatest amount of information is available when a set of *n* members is ungrouped. We start with all *n* members. Given that we have *n* items, we have *n* clusters, each containing just one member. The similarity between documents was computed in accordance with the frequency distribution of the letter patterns. The GPyM_TM software was used to carry out the clustering task. GPyM_TM is a Python package for performing clustering using the GSDMM model. The algorithm extracts the topical structures present within the corpus and clusters the documents with the potential to automatically identify the number of clusters. Clusters are then merged in successive iterations until one single cluster that contains all subsets is formed, as shown in Figure 5.

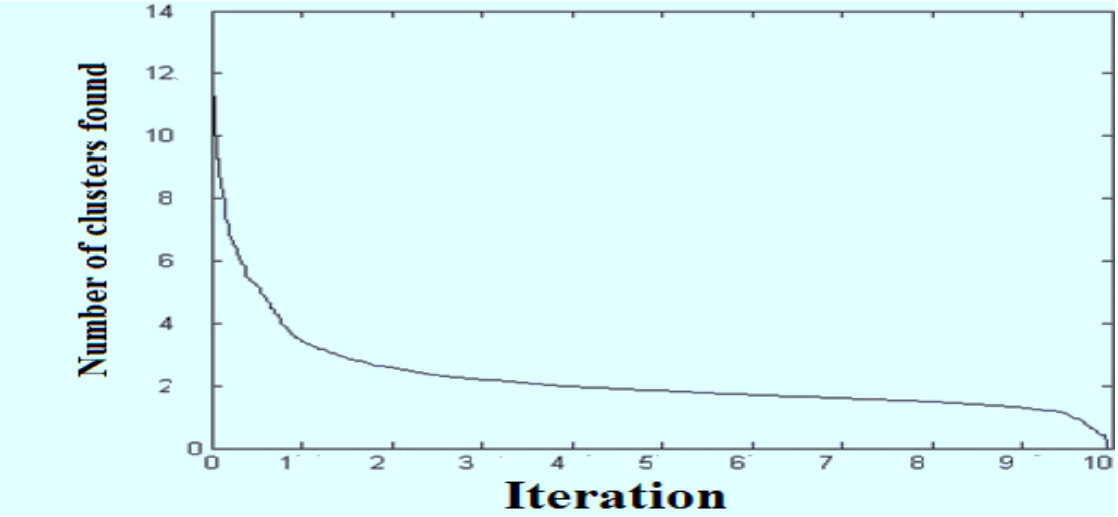

Figure 5. Number of clusters generated

Each lyric was assigned to a cluster through the probabilistic generative GSDMM model. The lyrics were clustered based on the mixture weights, such that the likelihood of documents/lyrics being related was calculated as the sum of the total probability overall mixture components in all the documents/lyrics. The documents/lyrics were assigned to clusters based on the highest conditional probability, as shown in Figure 5.

One advantage of this clustering approach is that it offers a solution to a traditional problem in cluster analysis—the decision concerning the optimal number of clusters that will fit a dataset. The strong tendency toward left-branching that is associated with other clustering methods is avoided with the GSDMM model, and thus so is the issue of identifying the optimal number of clusters.

4.5. Profile alignment

In our case, the lyrics were assigned to four distinct classes based on the distinctive features or variables of letter pairs. The clusters generated were compared to author profiles to evaluate the performance of the technique. This is technically known as the profile-based method. Clusters 1–6 were large and included the lyrics written by Morsi Gameel Aziz. Clusters 7–9 included the lyrics written by other poets. For convenience, clusters 1–6 were grouped together as they included lyrics written by the same author. It was decided to merge them under just one larger cluster, named X. In assigning the lyrics to clusters, it was clear that GSDMM offered the potential to infer and determine the number of clusters automatically, providing a good balance between completeness and homogeneity in the clustering results.

In terms of the disputed lyrics—Ya ma Alqamar 'ealbab Yama Anadyloh (O Ma, Moon-Like Sweetheart Is by the Door!), Hubak Nar (Your Love is Glowing in My Heart. Fire of Passion!), Btlwmwny Lyh law shwftom 'eynyh (Blame Me Not; Why on Earth Would You?, Seen, Not Told!), and Ana shwft Jamal (Ma, I Beheld the Beauty (of Nasser)!)—all were found to be grouped with the other lyrics clearly attributed to Aziz. Thus, the research finds that the disputed lyrics were all written by Aziz and not anyone else.

5. Discussion

The statistical findings generated through the clustering-based stylometric analysis indicate clearly that all the disputed lyrics were assigned to clusters with lyrics clearly attributed to Morsi Gameel

Aziz, suggesting that he is the real author of these lyrics. Although the study is limited to the questions of the authorship of Morsi Gameel Aziz's lyrics, the findings of the study can be extended to the authorship attribution of Arabic literature in general and lyrics in particular. The proposed approach can be usefully used in addressing many of the authorship issues and poetry theft cases that have come to the surface over the recent years. It can be concluded that the letter mapping and morphological patterns of each author can thus serve as distinctive linguistic stylometric features that can be usefully used for determining the authors of disputed texts in Arabic. It should be noted that letter mapping and morphological patterns have long been disregarded in authorship applications in Arabic which were traditionally based on stemming procedures as a pre-processing procedure in NLP applications of Arabic (Omar, Hamouda, 2020). In this regard, the distinctive morphological features of Arabic morphology should be considered in NLP applications including authorship attribution. Finally, the results of the study agree with the bulk of computational linguistics and literary computing studies, indicating that computer-based stylometry methods can usefully be implemented to enhance the performance of authorship attribution (Craig, 1999; Holmes, 1995b).

6. Conclusions

This study addressed the issue of the authorship of certain of Morsi Gameel Aziz's lyrics, first queried in the early 1960s and is still a controversial topic in Arab and Egyptian literary circles. To do so, a clustering-based stylometry approach was developed. The study employed a GSDMM model for dimensionality reduction and document clustering. It was obvious that the GSDMM model does not support counting tokens with multiplicity, which generally has little value in short text documents. It can thus be claimed that the GSDMM model can usefully be applied for the reduction of dimensionality in short texts in Arabic. The findings indicated that all the disputed lyrics were assigned to clusters with lyrics clearly attributed to Morsi Gameel Aziz, suggesting that he is the real author of these lyrics. The computational authorial analysis largely confirms the findings of some previous studies indicating that the disputed lyrics were written by Aziz. The findings of this study, however, are based on more solid foundations. These findings have their implications for authorship attribution in Arabic literature. Disputed texts in Arabic literature can be assigned to their real authors based on detecting stable linguistic patterns, yielding reliable performance in assigning authorship. Indeed, although the study focused on the lyrics of Morsi Gameel Aziz, the results of the study can be extended to Arabic lyrics and songs in general.

Acknowledgments

I take this opportunity to thank Prince Sattam Bin Abdulaziz University in Saudi Arabia alongside its Scientific Deanship, for all technical support it has unstintingly provided towards the fulfillment of the current research project.

References

Abu Rabiah, E. (2020). Lexical measures for testing progress in Hebrew as Arab students' L2. *Journal of Language and Linguistic Studies, 16*(3), 1096-1114.

Aggarwal, C. C., & Reddy, C. K. (2018). *Data Clustering: Algorithms and Applications*: CRC Press.

Al-Falahi, A., Ramdani, M., & Bellafkih, M. (2017). Machine learning for authorship attribution in Arabic poetry. *International Journal of Future Computer and Communication (IJFCC), 1*(6), 42-46.

Al-Falahi, A., Ramdani, M., & Bellafkih, M. (2019). Arabic Poetry Authorship Attribution using Machine Learning Techniques. *Journal of Computer Science, 15*(7), 1012.1021.

Amensisa, A. D., Patil, S., & Agrawal, P. (2018). *A survey on text document categorization using enhanced sentence vector space model and bi-gram text representation model based on novel fusion techniques.* Paper presented at the 2018 2nd International Conference on Inventive Systems and Control (ICISC).

Bagavandas, M., & Manimannan, G. (2008). Style Consistency and Authorship Attribution: A Statistical Investigation. *Journal of Quantitative Linguistics, 15*(1), 100-110. doi:10.1080/09296170701803426

Balossi, G. (2014). *A Corpus Linguistic Approach to Literary Language and Characterization: Virginia Woolf's The Waves*: John Benjamins Publishing Company.

Burges, C. J. C. (2010). *Dimension Reduction: A Guided Tour*: Now Publishers.

Burrows, J. (2002). 'Delta'—A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing, 17*(3), 267–287.

Burrows, J. (2007). All the Way Through: Testing for Authorship in Different Frequency Strata *Literary and Linguistic Computing, 22*(1), 27–47. doi:doi.org/10.1093/llc/fqi067

Coyotl-Morales, R. M., Villaseñor-Pineda, L., Montes-y-Gómez, M., & Rosso, P. (2006). Authorship Attribution Using Word Sequences. In C. O. J. A. Martínez-Trinidad J.F., Kittler J. (Ed.), *Progress in Pattern Recognition, Image Analysis and Applications, 4225*. Berlin, Heidelberg: Springer.

Crabb, P., Antonia, A., & Craig, H. (2014). Who wrote 'A Visit to the Western Goldfields'? Using Computers to Analyse Language in Historical Research. *History Australia, 11*(3), 177-193. doi:10.1080/14490854.2014.11668539

Craig, H. (1999). Authorial attribution and computational stylistics: If you can tell authors apart, have you learned anything about them? *Literary and Linguistic Computing, 14*, 103–113.

Craig, H., & Greatley-Hirsch, B. (2017). *Style, Computers, and Early Modern Drama: Beyond Authorship*. Cambridge Cambridge University Press.

Craig, H., & Kinney, A. F. (2009). *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge: Cambridge University Press.

Daelemans, W. (2013). Explanation in Computational Stylometry. In G. A. (Ed.), *Computational Linguistics and Intelligent Text Processing* (pp. 451-462). Berlin, Heidelberg: Springer.

Dauber Jr, E. G. (2020). *Stylometric Authorship Attribution Techniques and Analysis for Collaborative Platforms.* Drexel University.

Denis, D. J. (2020). *Univariate, Bivariate, and Multivariate Statistics Using R: Quantitative Tools for Data Analysis and Data Science*: Wiley.

Eder, M., Piasecki, M., & Walkowiak, T. (2017). An open stylometric system based on multilevel text analysis. *Cognitive Studies [Études cognitives], 17*. doi:doi.org/10.11649/cs.1430

El Bakly, A. H., Darwish, N. R., & Hefny, H. A. A Survey on Authorship Attribution Issues of Arabic Text.

Everitt, B. (2009). *Multivariable Modeling and Multivariate Analysis for the Behavioral Sciences*: CRC Press.

Everitt, B., & Hothorn, T. (2011). *An Introduction to Applied Multivariate Analysis with R*: Springer New York.

Gan, G., Ma, C., & Wu, J. (2020). *Data Clustering: Theory, Algorithms, and Applications, Second Edition*: SIAM.

Gómez-Adorno, H., Posadas-Duran, J.-P., Ríos-Toledo, G., Sidorov, G., & Sierra, G. (2018). Stylometry-based Approach for Detecting Writing Style Changes in Literary Texts. *Computación y Sistemas, 22*(1), 47-53. doi:doi.org/10.13053/cys-22-1-2882

Hennig, C., Meila, M., Murtagh, F., & Rocci, R. (2015). *Handbook of Cluster Analysis*: CRC Press.

Holmes, D. (1995a). Authorship Attribution. *Computers and the Humanities, 28*, 87-106.

Holmes, D. (1995b). The Federalist revisited: new directions in autorship attribution. *Literary and Linguistic Computing, 10*, 111–127.

Holmes, D. (1998). The Evolution of Stylometry in Humanities Scholarship. *Literary and Linguistic Computing, 13*(3), 111-117. doi:doi:10.1093/llc/13.3.111

Hoover, D. L. (2003). Multivariate analysis and the study of style variation. *Literary and Linguistic Computing, 18*(4), 341–360. doi:doi.org/10.1093/llc/18.4.34

Hoover, D. L. (2004). Testing Burrows' delta. *Literary and Linguistic Computing, 19*(4), 453–475.

Hoover, D. L., Culpeper, J., & O'Halloran, K. (2014). *Digital Literary Studies: Corpus Approaches to Poetry, Prose, and Drama*. London; New York: Routledge.

Hussein, T. (1927). *On the Pre-Islamic Poetry* (2nd ed.). Cairo, Egypt: Hindawi Foundation for Education and Culture.

Iqbal, F., Debbabi, M., & Fung, B. C. M. (2020). *Machine Learning for Authorship Attribution and Cyber Forensics*: Springer International Publishing.

Ison, D. C. (2020). Detection of Online Contract Cheating through Stylometry: A Pilot Study. *Online Learning, 24*(2), 142-165.

Jackson, J. E. (2005). *A User's Guide to Principal Components*: Wiley.

Jockers, M. L. (2014). *Text Analysis with R for Students of Literature*: Springer International Publishing.

Jolliffe, I. T. (2006). *Principal Component Analysis*: Springer New York.

Juola, P. (2008). *Authorship Attribution*: Published, sold, and distributed by now Publishers.

Lagutina, K., Boychuk, E., Vorontsova, I., & Paramonov, l. (2019). *A Survey on Stylometric Text Features*. Paper presented at the 25th Conference of Open Innovations Association (FRUCT), Helsinki, Finland.

Linwei Li, Guo, L., He, Z., Jing, Y., & Wang, S. (2019). X-DMM: Fast and Scalable Model Based Text Clustering. *The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*, 4197-4204.

Love, H. (2002). *Attributing Authorship: An Introduction*. Cambridge: Cambridge University Press.

Mathews, R. A., & Merriam, T. V. (1993). Neural Computation in Stylometry I: An Applicationto the Works of Shakespeare and Fletcher. *Literary and Linguistic Computing, 8*(4), 203-209.

Mazarura, J., & Waal, A. d. (2016, 30 Nov.-2 Dec. 2016). *A comparison of the performance of latent Dirichlet allocation and the Dirichlet multinomial mixture model on short text.* Paper presented at the 2016 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech).

Moisl, H. (2008). Data Normalization for Variation in Document Length in Exploratory Multivariate Analysis of Text Corpora. *INFOS 2008: The 6th International Conference on Informatics and Systems Special Track On Natural Language Processing*, 85-92.

Moisl, H. (2015). *Cluster Analysis for Corpus Linguistics*: De Gruyter.

Moisl, H., & Maguire, W. (2008). Identifying the Main Determinants of Phonetic Variation in the Newcastle Electronic Corpus of Tyneside English. *Journal of Quantitative Linguistics, 15*(1), 46-69. doi:10.1080/09296170701794302

Omar, A. (2021). Identifying Themes in Fiction: A Centroid-Based Lexical Clustering Approach. *Journal of Language and Linguistic Studies, 17*(Special Issue 1), 580-594.

Omar, A., & Aldawsari, B. D. (2019). Towards a Linguistic Stylometric Model for the Authorship Detection in Cybercrime Investigations. *International Journal of English Linguistics, 9*(5), 182-192.

Omar, A., Elghayesh, B. I., & Kassem, M. (2019). Authorship Attribution Revisited: The Problem of Flash Fiction A morphological-based Linguistic Stylometry Approach. *Arab World English Journal (AWEJ), 10*(3), 318-329.

Omar, A., & Hamouda, W. I. (2020). The Effectiveness of Stemming in the Stylometric Authorship Attribution in Arabic. *International Journal of Advanced Computer Science and Applications(IJACSA), 11*(1), 116-121. doi:10.14569/IJACSA.2020.0110114

Pask, K. (2002). Plagiarism and the Originality of National Literature: Gerard Langbaine. *ELH, 69*(3), 727-747.

Pyle, D. (1999). *Data Preparation for Data Mining*: Elsevier Science.

Roelleke, T. (2013). *Information Retrieval Models: Foundations and Relationships*: Morgan & Claypool Publishers.

Rudman, J. (1997). The State of Authorship Attribution Studies: Some Problems and Solutions. *Computers and the Humanities, 31*(4), 351-365.

Rudman, J. (2012). The State of Non-Traditional Authorship Attribution Studies—2012: Some Problems and Solutions. *English Studies, 93*(3), 259-274. doi:10.1080/0013838X.2012.668785

Savoy, J. (2012). Authorship Attribution: A Comparative Study of Three Text Corpora and Three Languages. *Journal of Quantitative Linguistics, 19*(2), 132-161. doi:10.1080/09296174.2012.659003

Schreibman, S., Siemens, R., & Unsworth, J. (2016). *A New Companion to Digital Humanities* (2nd ed.): Wiley-Blackwell.

Smith, M. W. A. (1992). Shakespeare, Stylometry and "Sir Thomas More'". *Studies in Philology, 89*(4), 434-444.

Srivastava, A. N., & Sahami, M. (2009). *Text Mining: Classification, Clustering, and Applications*: Taylor & Francis.

Stamatatos, E., Tschuggnall, M., Verhoeven, B., Daelemans, W., Specht, G., Stein, B., & Potthast, M. (2016). *Clustering by Authorship Within and Across Documents*. Paper presented at the CLEF 2016, Évora, Portugal.

Strome, E. (2013). "Raked from the Rubbish": Stylometric Authorship Attribution and the 1795 American Philosophical Society Education Contest. In J. B. (Ed.), *The Founding Fathers, Education, and "The Great Contest"* (pp. 45-65). New York: Palgrave Macmillan.

Timm, N. H. (2007). *Applied Multivariate Analysis*: Springer New York.

Varela, P. J., Albonico, M., Justino, E. J. R., & de Assis, J. L. V. (2020). Authorship Attribution in Latin Languages using Stylometry. *IEEE Latin America Transactions, 18*(04), 729-735.

Watt, J., Borhani, R., & Katsaggelos, A. K. (2020). *Machine Learning Refined: Foundations, Algorithms, and Applications*: Cambridge University Press.

Weiss, S. M., Indurkhya, N., Zhang, T., & Damerau, F. (2010). *Text Mining: Predictive Methods for Analyzing Unstructured Information*: Springer New York.

Wu, W., Xiong, H., & Shekhar, S. (2013). *Clustering and Information Retrieval*: Springer US.

Yang, S., Haung, G., & Cai, B. (2019). Discovering Topic Representative Terms for Short Text Clustering. *IEEE Access, 7*, 92037-92047. doi:10.1109/ACCESS.2019.2927345

Yin, J., & Wang, J. (2014). A Dirichlet Multinomial Mixture Model-based Approach for Short Text Clustering. *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 233–242. doi:doi.org/10.1145/2623330.2623715

Zhao, S., Li, S., Qi, L., & Da Xu, L. (2020). Computational intelligence enabled cybersecurity for the internet of things. *IEEE Transactions on Emerging Topics in Computational Intelligence, 4*(5), 666-674.

Zheng, L., & Zheng, H. (2020). Authorship Attribution via Coupon-Collector-Type Indices. *Journal of Quantitative Linguistics, 27*(4), 321-333. doi:10.1080/09296174.2019.1577939

Zhiguo, G., Luo, X., Chen, J., Wang, F. L., & Lei, J. (2011). *Emerging Research in Web Information Systems and Mining: International Conference, WISM 2011, Taiyuan, China, September 23-25, 2011. Proceedings*: Springer Berlin Heidelberg.

Zhu, H., Lei, L., & Craig, H. (2020). Prose, Verse and Authorship in Dream of the Red Chamber: A Stylometric Analysis. *Journal of Quantitative Linguistics*, 1-17. doi:10.1080/09296174.2020.1724677

AUTHOR BIODATA

Abdulfattah Omar is an Associate Professor of English Language and Linguistics in the Department of English, College of Science & Humanities, Prince Sattam Bin Abdulaziz University (KSA). Also, he is a standing lecturer of English Language and Linguistics in the Department of English, Faculty of Arts, Port Said University, Egypt. Dr. Omar received his PhD degree in computational linguistics in 2010 from Newcastle University, UK. His research interests include computational linguistics, digital humanities, discourse analysis, and translation studies.