The Effectiveness of Stemming in the Stylometric Authorship Attribution in Arabic

Abdulfattah Omar*[1]
College of Science and Humanities
Prince Sattam Bin Abdulaziz University, Saudi Arabia
Department of English, Faculty of Arts, Port Said University

Wafya Ibrahim Hamouda[2]
Department of Foreign Languages
Faculty of Education, Tanta University
Egypt

***Abstract*—The recent years have witnessed the development of numerous approaches to authorship attribution including statistical and linguistic methods. Stylometric authorship attribution, however, remains among the most widely used due to its accuracy and effectiveness. Nevertheless, many authorship problems remain unresolved in terms of Arabic. This can be attributed to different factors including linguistic peculiarities that are not usually considered in standard authorship systems. In the case of Arabic, the morphological features carry unique stylistic features that can be usefully used in testing authorship in controversial texts and writings. The hypothesis is that much of these morphological features are lost due to the execution of stemming. As such, this study is concerned with investigating the effectiveness of stemming in the stylometric applications to authorship attribution in Arabic. In so doing, three Arabic stemmers GOLD stemmer, Khoga stemmer, Light 10 stemmer are used. By way of illustration, a corpus of 2400 news articles written by different 97 authors is designed. To evaluate the effectiveness of stemming, the selected articles (both stemmed and unstemmed texts) are clustered using cluster analysis methods. Comparisons are made between clustering structures based on stemmed and unstemmed datasets. The results indicate that stemming has negative impacts on the accuracy of the clustering performance and thus on the reliability of stylometric authorship testing in Arabic. The peculiar stylistic features of the affixation processes in Arabic can, thus, be usefully used for improving the performance of authorship attribution applications in Arabic. It can be finally concluded that stemming is not effective in the stylometric authorship applications in Arabic.**

***Keywords*—Authorship attribution; cluster analysis; GOLD stemmer; Khoga stemmer; Light 10 stemmer; stemming; stylometry**

I. Introduction

The recent years have witnessed an increasing use of stylometric approaches in addressing different authorship problems. These have been mainly based on the investigation of the lexical (e.g. frequency of distinctive words, discourse markers, and modal verbs) and structural (e.g. use of chunks, type of sentence, and sentence length) properties of the texts as a clue for identifying authors of controversial texts. In spite of the success of these approaches in solving different authorship problems of various controversial documents, presently, they are ineffective and thus, unreliable in addressing authorship problems in Arabic. This can be attributed to the peculiar linguistic features of Arabic. As thus, this study investigates the effectiveness of stemming in the stylometric authorship attribution in Arabic. The hypothesis is that derivational and inflectional morphemes (which are normally removed in stemming applications) carry unique stylistic features that may be useful in identifying authors of controversial texts. By way of illustration, this study is based on a corpus of 2400 news articles written by 97 authors derived from four newspapers. Three stemming algorithms were used. To evaluate the effectiveness of stemming, datasets (both stemmed and unstemmed texts) were clustered using hierarchical cluster analysis methods. The remainder of this article is organized as follows. Section 2 asks the research question concerning the effectiveness of stemming in the stylometric authorship applications in Arabic. Section 3 surveys the authorship attribution literature and the emergence of stylometric approaches in authorship studies. Section 4 defines the methods and procedures. Section 5 is an experimental analysis of the effect of stemming on the accuracy and reliability of text clustering performance. Section 6 is conclusion.

II. Research Question

Stemming can be broadly defined as the practice of conflating semantically equivalent word variants into the same root by removing derivational and inflectional affixes [1-3]. Technically speaking, it is a procedure that tries to remove inflectional and derivational suffixes to conflate word variants into the same stem or root [4]. The basic concept of stemming is that words of identical stem or root that refer to the same concept must, therefore, be grouped under the same type. Paice makes it clear that the function of stemming is to conflate all words which share equivalent semantics and share identical stems [5].

Many stemmers have been understood and through this, developed for a colossal range of languages including English, French, German, Dutch, Swedish, Latin, Malay, Indonesian, Slovene, Turkish, Arabic, and Hebrew. Leah, Lisa, et al. [6] point out that stemmers tend to be bespoke and exclusive to each independent language [6]. Building stemmers accordingly requires some linguistic knowledge of the language and an understanding of the needs of information retrieval. The concept of all stemmers is the reduction of the corpora size so that data mining processes (e.g. Information Retrieval, text clustering, etc.) systems work faster and more effectively. All stemmers have one thing in common: they are all designed to remove derivational and inflectional affixes and conflate word variants of the same base into a common

Paper Submission Date: December 25, 2019
Acceptance Notification Date: January 14, 2020
*Corresponding Author

term. Some stemmers are designed for both derivational and inflectional affixes; others are designed for only suffixes, and some are designed solely for handling simple plurals. In English, just like many Western European languages, stemming is predominantly a methodology of removing a suffix. That is, stemming is a procedure for removing suffixes attached at the end of words. The point is that stemming algorithms for English and other European languages normally fail to consider prefixes and infixes. In this, English stemming is primarily concerned with the morphology of suffixes.

With regard to Arabic, two main approaches have been developed. These are the root-based method and the light stemming approach. Elrajubi [7] explains that light stemming algorithms have the purpose to only remove prefixes and suffixes from the words, whereas root-based algorithms eliminate prefixes, suffixes and infixes. Due to the numerous problems caused by root-based algorithms, there is a tendency to use light stemmers. Light stemmers are more concerned with removing the prefix and suffix of a word [8]. Obvious examples are Khoga stemmer, Light 10 stemmer, and GOLD stemmer. In spite of the increasing use of stemming as a requirement or a pre-processing step in different NLP applications, there is no stemming algorithm that is 100% precise. To address this problem, dissimilar studies have been recently focussed on evaluating and comparing the performance of Arabic stemmers to provide users and researchers with answers about the most appropriate algorithm for their tasks [7, 9, 10]. Nevertheless, there are no definite answers to the effectiveness of stemming in stylometric authorship applications in Arabic. In the face of this problem, this study asks this research question: What is the effectiveness of stemming in stylometric authorship applications in Arabic?

III. Literature Review

Authorship attribution, also known as authorship recognition, is the process of looking for salient features in a piece of writing that relates the work to its author. Craig [11] points out that authorship studies have objectives of 'yes or no' declarations to present problems, and are said to avoid observable features if possible; due to operating at the base strata of language, where imitation or deliberate variation can be rejected [11]. The idea of authorship attribution is very old. Love [12] mentions that it ventures back to the period of the well-renowned library of Alexandria and accordingly, comprise the construction of the "Jewish and Christian biblical canons". The motive behind authorship attribution studies is that many works were written anonymously, and many others raise suspicion about their real author, and ultimately, historical evidence is sparse or indeed lacking. Traditionally, work on authorship attribution was conceived as an organized scholarly enterprise where it was not the achievements of an expert or scholar in authorship, but the contributions of a scholar to which the fortitude of authorship had constantly been a vital constituent into other investigative natures [12]. There are many examples where the task of identifying the author of a particular document was the job of politicians, journalists, and lawyers [13, 14]. Studies in this tradition often used criteria for relating works to authors on chronological and epistemological bases. One problem with such methods is that it [12] is often difficult to find reliable historical facts or knowledge-based evidence that will help in the identification of authors. Furthermore, these studies were based on what can be considered philological approaches making no use of replicable methods and thus the results were not objective and thus unreliable.

In the face of these limitations, empirically-driven approaches for authorship attribution problems were developed. The claim was that authorship attribution applications should be algorithmically processed without any reference to existing analytical results or personal knowledge of authors [15]. The mainstream of these approaches is described in the literature as stylometry. Stylometry is a quantitative inquiry into the individualities of an author's style and technique. Laan [16] defines the term as a technique with a purpose to maintain the comprehension of the, often elusive, characteristics and individualities of an author's style, or at least a fraction of it, through enumerating some of its features and qualities [16]. Merriam and Matthews [17] claiming "stylometry attempts to capture quantitatively the essence of an individual's use of language" [17]. Stylometric studies have been mainly based on computational and quantitative methods in order to reach solid conclusions regarding the authorship of a given text [18]. Accordingly, numerous studies have come to give empirical solutions to different controversial authorship issues using quantitative methods for investigating the stylistic and linguistic properties of authors.

One of the pioneering examples of the use of stylometric analysis in authorship problems is Mosteller and Wallace [19] attempt to give internal evidence for the authors of the dubious Federalist Papers based on linguistic and stylistic properties of the authors. These are 77 Federalist Papers written during 1787-1788 to Alexander Hamilton, John Jay and James Madison. These papers were published in newspapers under the pseudonym of Publius until they were collected with eight more articles to form a volume. There was a consensus about the authorship of these Papers that John Jay had authored 5 papers in the volume; while Hamilton authored 51 papers; Madison wrote 14 and both Madison and Hamilton co-authored three. The authorship of 12 papers in the volume was rather disputed since it was difficult to find out which of the two, Madison or Hamilton had authored those Papers [20, 21]. On their part, Mosteller and Wallace [19], employed tools of statistical analysis in order to investigate the mystery of authorship of the Federalist papers in the early 1960s, using function words as discriminators. The objectivity and replicability of the proposed approach opened the way to the computerized age of authorship attribution.

The basic assumption behind stylometric testing of authorship attribution is that, Holmes [22] contends, authors have unawareness of their characteristic styles. These are styles which cannot deliberately be influenced yet acquire features which are reckonable and are thus, highly unique [22] and the identification of such personal distinctive linguistic and stylistic features makes it possible to detect an author's signature and distinguish the writing of one author from another or others. In this way, researchers and particularly the statisticians, Knaap and Grootjen [23] had a tendency to

investigate the lexical features of texts in order to make predictions about possible authors. As thus, the search for the most frequent words has been one of the most widely used methods for determining the author of a given work [24, 25]. Garcia and Martin [26] explain that statisticians attempted over the last decade to solve some controversial authorship problems by finding a formula grounded on the computation of tokens, word-types, and most frequently-used words. They contend that computational statisticians have tended to investigate, what they call, the 'Lexical Richness' of authors in order to propose a reliable approach to authorship attribution. On the other hand, Morton [27] argued that the use of rare words is a good indication for determining the author of a given text as this enables one writer to be prominent from another. He explains that occurring words communicate a multitude of essentials, which acquired the belief to demonstrate brilliance in writing. These were noted as "the range of a writer's interests, the precision of his observation and the imaginative power of his comparisons", and thus, exhibit his command of pattern and of interchanges [27]. Similarly, Blatt [28] asserts that rare words are quite noticeable and can be considered the writer's favorite words which makes it easier and accurate to use them as an indicator for determining authors.

The ineffectiveness of the lexical representation of texts in resolving different authorship problems, however, has led to the development of new methods. The lexical representation of texts has come to be known today as the traditional way of doing authorship attribution. It has been criticized for its ineffectiveness in providing solutions for the practical applications of authorship attribution [29, 30]. The claim is that isolated or single words are not enough for assigning disputed texts to their possible writers. The idea is simply that single words are not enough to capture the structure of documents. Different studies, therefore, have been more concerned with the morphological, syntactic, and structural features of texts (e.g. morphologically complex words, use of function words, sentence length, compounding, and punctuation).

In spite of the reasonable success of the stylometric methodologies in providing answers for many authorship problems, verifying the authorship of Arabic texts still represents a real challenge for the practical applications of author identification. This may be due to the fact that very few studies have been concerned with authorship attribution in Arabic, in which differences in language systems represent further challenges. This study tends to address this gap in literature by investigating the effectiveness of stemming in the stylometric authorship attribution in Arabic.

IV. Methods and Procedures

To evaluate the effectiveness of stemming in stylometric authorship attribution, three Arabic stemmers were used. These are Khoga stemmer, Light 10 stemmer and GOLD [31]. The rationale is that these stemmers are fast and straightforward algorithms and they are widely used in different NLP applications including information retrieval (IR) and document clustering. Although different studies have pointed to the idea that Light 10 outperforms many other stemmers, it was thought that it would be appropriate to include different stemmers for validity purposes.

This study is based on a corpus of 2400 news articles written by 97 authors derived from three newspapers. These are Al-Ahram, Asharq Al-Awsat, and Al-Hayat. Articles covered the period between 2016 and 2018. All selected articles are written in Modern Standard Arabic (MSA). This is an overly formal version of Arabic and differs substantially from spoken dialects. The rationale is that in MSA, core grammar and vocabulary remain constant. The root-and-pattern system is almost the same in all MSA dialects.

For assigning texts to their authors, cluster analysis methods were used. Cluster analysis is widely acknowledged as a successful technique for organizing any unorganized sets of documents [32]. It is an exploratory multivariate technique for systematically finding relatively homogeneous clusters of cases based on proximity measures without prior assumptions about differences within sets of data investigated [33-35]. It is a deterministic process that identifies discrete categories under any inherent structure in the data [36-41]. It is thus an inductive technique that explicitly attempts to group data sets into discrete classes [42, 43]. The aim of cluster analysis can be summarised as grouping a collection of objects into subsets where members of each subgroup are more closely related to one another than members assigned to the other group/s. Groups are technically referred to as clusters. Given a corpus of 2400 documents, these can be clustered where members of each cluster share specific characteristics. In authorship recognition applications, the assumption is that texts grouped together are more likely to be written by the same author. To perform cluster analysis, Euclidian distance, being a straightforward measure, was used. Euclidian distance is the most widely used and is reported to provide reliable results in general. As for the clustering method, Ward linkage is used. The rationale is that the Ward linkage clustering (or what is usually referred to as increase in sum of squares) with Euclidean measure seems to be the most convenient for the present case because it makes the clearest partitioning of the matrix rows.

V. Analysis

In order to evaluate the effectiveness of stemming in stylometric authorship applications, two processes were carried out. First, similar texts were grouped together assuming that texts grouped together are more likely to be written by the same author. Second, clustering structures were compared to the bibliographic information of each author. In order to compute the similarity between texts and group similar texts together, the Ward linkage clustering method with Euclidean distance measure was used. As a result, the matrix rows are assigned to clearly identified four groups. One advantage of this clustering is that it offers a solution for a traditional problem in cluster analysis-the decision of the optimal number of clusters that fits a dataset. The strong tendency towards left-branching that is associated with other clustering methods is avoided with Ward clustering. The matrix rows are assigned to four main groups as shown in Fig. 1.

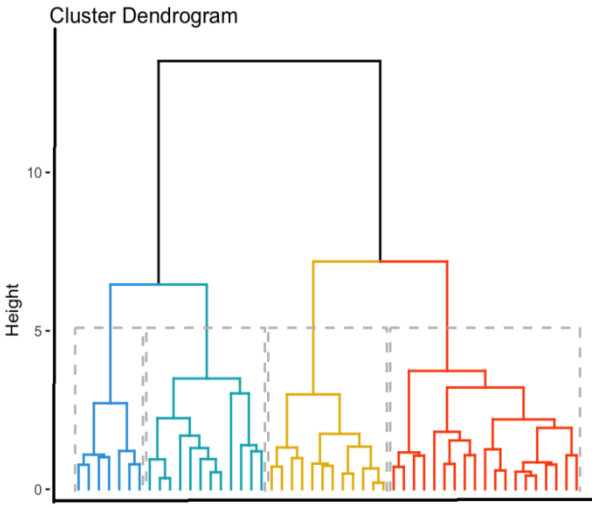

Fig. 1. Hierarchical Cluster Analysis of the Data Matrix.

For clustering validity purposes, two approaches were used. These were cross-validation and relative comparison. The purpose was to validate the foregoing analysis by seeing whether the same analytical methods applied to an alternative representation of the data give identical or at least similar results. In a cross-validation approach, the texts were randomly divided into two subsets, say *A* and *B*, and the cluster analysis is carried out separately on each of *A* and *B*. The similarity of the results is the indication of validity [44]. Comparison shows a close fit between the results as there is a total correspondence between the structures based on the data matrix composed of all the 2400 rows and the structures based on the random distribution of these 2400 rows into two groups.

For relative comparison analysis, a relative approach was based on comparing the clustering structure, generated by the same algorithms but using an alternative representation of the data; this was done by cluster analyzing a principal component reduction of the data matrix. Analysis showed that there is a close fit between the two clustering structures despite the minor differences. Consequently, it can be claimed that the agreement between the clustering structures supports the validity of hierarchical cluster analysis results. As a final step, the clustering structure obtained here was compared to the bibliographic information of each author. Members of each class or cluster were compared to the stories of each author.

The analysis was carried out four times and yielded specific results, namely, that the Light 10 dataset brought a 67% accuracy rate. The Khoga stemmer dataset yielded a 64% accuracy rate, whilst the GOLD and "Without stemming" datasets yielded accuracy rates of 61% and 78%, respectively.

The findings indicate clearly that clustering performance works much better without executing stemming. This can be attributed to two reasons. First, stemmers commit numerous errors. The three stemmers merged words that are different in form and are also semantically distinct and different from each other. Furthermore, stemmers find no solutions to homographs. This means that stemmers conflate word forms that are completely different in meaning. It was found out that the stemmers made two kinds of error: over-stemming and under-stemming. Over-stemming refers to forming larger stem classes where unrelated forms are wrongly conflated. Under-stemming, on the other hand, refers to failing to conflate variant forms of the same stem leaving them ungrouped.

It was found out that clustering structures based on Gold stemmer were the least reliable. This is due to the so many problems associated with this stemmer. These problems can be summarized as follows. First, it removes only one suffix from a word, due to its nature as a single pass algorithm. Second, it fails to form words from the stems, or matches the stems of like meaning words. Finally, its large set of rules and the recoding stage affect the speed of execution. The Khoga stemmer comes second in terms of the effectiveness. One good advantage of this stemmer attempts to find solutions to irregularities or what the compilers call non-formulaic changes (i. e. irregular plurals) by providing a lexicon within the stemmer. The problem with such non-formulaic changes is that they are unpredictable and stemming without the usage of a lexicon is fundamentally unmanageable without presenting errors.

Based on the quantitative results of the performance of each stemmer, Light 10 Stemmer can be claimed to be the most effective stemmer for Arabic data in relation to stylometric applications. This stemming algorithm is a procedure for removing the derivational and inflectional suffixes from Arabic words. However, putting the algorithm into practice and test, it is observed that it has some shortcomings. First, it makes the two kinds of errors of over-stemming and under-stemming. That is, it is sometimes too aggressive in conflation and groups words like execute and executive together and sometimes it is too weak and misses words so that they are not conflated. Second, it articulates terms (stems) that are not words and are too difficult to identify. Finally, it ignores prefix removal completely.

The second reason that can be attributed to the poor performance of the text clustering based on stemmed entries is the peculiar nature of the morphological system in Arabic. With stemming, the morphological features of Arabic, which carry unique stylistic features that distinguish authors, are lost to a great extent.

VI. Conclusion

This study addressed the question of the low performance of stylometric authorship applications of Arabic texts. The hypothesis was that part of the problem is related to the lack of consideration to linguistic peculiarities that are not usually considered in standard authorship systems. In light of this argument, the study investigated the effectiveness of stemming which is a prerequisite in different stylometric authorship systems on the performance of stylometric authorship systems in Arabic. The results indicate clearly that morphological information can be usefully used for improving the performance of authorship attribution and detection in Arabic texts due to the unique stylistic features of the affixation processes in Arabic. Controversial texts in Arabic can, thus, be assigned to their authors based on detecting stable morphological patterns with reliable authorship performance. Although the proposed system was tested only

on literary texts written in Standard Arabic, the implications of the study can be practically used for the authorship problems in other text genres including emails, newsgroup messages, Facebook posts, and tweets as well as different Arabic varieties which still represent a real challenge for the practical applications of author identification.

ACKNOWLEDGMENT

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

[1] L. C. Jain, H. S. Behera, J. K. Mandal, and D. P. Mohapatra, Computational Intelligence in Data Mining - Volume 3: Proceedings of the International Conference on CIDM, 20-21 December 2014. Springer India, 2014.

[2] M. Bramer, Principles of Data Mining. Springer London, 2007.

[3] R. Feldman, J. Sanger, and C. U. Press, The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press, 2007.

[4] J. Savoy, "A stemming procedure and stopword list for general French corpora," J. Am. Soc. Inf. Sci., vol. 50 no. 10, pp. 944-952, 1999.

[5] C. D. Paice, "Method for evaluation of stemming algorithms based on error counting," J. Am. Soc. Inf. Sci., vol. 47, no. 8, pp. 632-649, 1996.

[6] S. L. Leah, B. Lisa, and E. C. Margaret, "Improving stemming for Arabic information retrieval: light stemming and co-occurrence analysis," Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 275-282, 2002.

[7] O. M. Elrajubi, "An improved Arabic light stemmer," in 2013 International Conference on Research and Innovation in Information Systems (ICRIIS), 2013, pp. 33-38.

[8] L. S. Larkey, L. Ballesteros, and M. E. Connell, "Light Stemming for Arabic Information Retrieval," in Arabic Computational Morphology: Knowledge-based and Empirical Methods, A. Soudi, A. v. d. Bosch, and G. Neumann, Eds. Dordrecht: Springer Netherlands, 2007, pp. 221-243.

[9] Y. Jaafar, D. Namly, K. Bouzoubaa, and A. Yousfi, "Enhancing Arabic stemming process using resources and benchmarking tools," Journal of King Saud University - Computer and Information Sciences, vol. 29, no. 2, pp. 164-170, 2017/04/01/ 2017.

[10] Y. Al-Lahham, K. Matarneh, and M. Hassan, "Conditional Arabic Light Stemmer: CondLight," International Arab Journal of Information Technology, vol. 17, 12/05 2018.

[11] H. Craig, "Stylistic Analysis and Authorship Studies," in A Companion to Digital Humanities, S. Schreibman, R. Siemens, and J. Unsworth, Eds. Oxford: Blackwell, 2004, pp. 273-288.

[12] H. Love, Attributing Authorship: An Introduction. Cambridge: Cambridge University Press, 2002, p. 240p.

[13] P. Juola, J. Sofko, and P. Brennan, "A Prototype for Authorship Attribution Studies," Lit Linguist Computing, vol. 21 (2), no. 2, pp. 169-178, June 1, 2006 2006.

[14] P. Juola, "Authorship Attribution," Foundations and Trends R in Information Retrieval, vol. 1, no. 3, pp. 233-334, 2008.

[15] H. Moisl, "Using electronic corpora in historical dialectology research," in Studies in English and European Historical Dialectology, M. Dossena and R. Lass, Eds. Brussels; Frankfurt: Peter Lang, 2009, pp. 68-90.

[16] N. M. Laan, "Stylometry and Method. The Case of Euripides," Lit Linguist Computing, vol. 10, no. 4, pp. 271-278, November 1, 1995 1995.

[17] T. Merriam and R. Matthews, "Neural Computation in Stylometry II: An Application to the Works of Shakespeare and Marlowe," Literary and Linguistic Computing, vol. 9, no. 1, pp. 1-6, 1994.

[18] G. Tambouratzis and M. Vassiliou, "Employing Thematic Variables for Enhancing Classification Accuracy Within Author Discrimination Experiments," Lit Linguist Computing, vol. 22, no. 2, pp. 207-224, June 1, 2007 2007.

[19] F. Mosteller and D. L. Wallace, Inference and Disputed Authorship: The Federalist (Addison-Wesley series in behavioral sciences. Quantitative methods.). Reading, Mass.: Addison-Wesley Pub. Co., 1964, pp. xv, 287 p.

[20] J. Rudman, "The Twelve Disputed 'Federalist' Papers: A Case for Collaboration," Proceedings Digital Humanities, pp. 353–356, 2012.

[21] J. Savoy, "The Federalist Papers revisited: A collaborative attribution scheme," in Proceedings of the Association for Information Science and Technology, Montreal, Quebec, Canada., 2013.

[22] D. Holmes, "The Evolution of Stylometry in Humanities Scholarship," Literary and Linguistic Computing vol. 13, pp. 111-117, 1998.

[23] L. v. d. Knaap and F. A. Grootjen, " Author identification in chatlogs using formal concept analysis," presented at the Proceedings of the 19th Belgian-Dutch Conference on Artificial Intelligence (BNAIC2007), Utrecht, The Netherlands, November 2007, 2007.

[24] J. F. Burrows, "Questions of Authorship: Attribution and Beyond A Lecture Delivered on the Occasion of the Roberto Busa Award ACH-ALLC 2001, New York," Computers and the Humanities, vol. 37, no. 1, pp. 5-32, 2003.

[25] J. F. Burrows, "All the Way Through: Testing for Authorship in Different Frequency Strata," Lit Linguist Computing, vol. 22, no. 1, pp. 27-47, April 1, 2007 2007.

[26] A. M. Garcia and J. C. Martin, "Function Words in Authorship Attribution Studies," Lit Linguist Computing, vol. 22, no. 1, pp. 49-66, April 1, 2007 2007.

[27] A. Q. Morton, "Once. A Test of Authorship Based on Words which are not Repeated in the Sample," Lit Linguist Computing, vol. 1, no. 1, pp. 1-8, January 1, 1986 1986.

[28] B. Blatt, Nabokov's Favorite Word Is Mauve: What the Numbers Reveal About the Classics, Bestsellers, and Our Own Writin. Simon & Schuster, 2017.

[29] E. Stamatatos, "A Survey of Modern Authorship Attribution Methods," Journal of American Society for Information Science and Technology, vol. 60, no. 3, pp. 538-556, 2009.

[30] M. S. Tamboli and R. S. Prasad, "Authorship Analysis and Identification Techniques: A Review," International Journal of Computer Applications, vol. 77, no. 16, pp. 11-15 2013.

[31] M. Rogati, S. McCarley, and Y. Yang, "Unsupervised learning of Arabic stemming using a parallel corpus," in Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, Stroudsburg, USA, 2003, vol. 1.

[32] H. Moisl, Cluster Analysis for Corpus Linguistics. Berlin, Munich, Boston: Walter de Gruyter, 2015.

[33] L. Kaufman and P. J. Rousseeuw, Finding Groups in Data: An Introduction to Cluster Analysis. Hoboken, New Jersey: John Wiley& Sons, INC, 1990.

[34] A. Fielding, Cluster and Classification Techniques for the Biosciences. Cambridge, UK ; New York: Cambridge University Press, 2007, pp. xii, 246 p.

[35] C. D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval. Cambridge: Cambridge University Press, 2008.

[36] G. W. Milligan, "Clustering Validation: Results and Implications for Applied Analyses," in Classification and Clustering, P. Arabie, Hubert, L.J. and De Soete, G, Ed. River Edge, NJ: World Scientific Publishing Co Pte Ltd, 1996.

[37] B. Everitt, S. Landau, and M. Leese, Cluster analysis, 4th ed. / Brian S. Everitt, Sabine Landau, Morven Leese. ed. London: Arnold ; New York : Oxford University Press, 2001, pp. viii, 237 p.

[38] G. Punj and D. W. Stewart, "Cluster Analysis in Marketing Research: Review and Suggestions for Application," Journal of Marketing Research, vol. 20, no. 2, pp. 134-148, 1983.

[39] J. F. Hair, Multivariate data analysis, 6th ed. ed. Upper Saddle River, N.J. ;London: Prentice Hall PTR, 2006, pp. xxiv, 899 p.

[40] M. R. Anderberg, Cluster analysis for applications (Probability and mathematical statistics ; 19). New York ; London: Academic Press, 1973, pp. xiii,359p.

[41] B. Everitt, Cluster analysis, 3rd ed. London: E. Arnold, 1993, pp. viii, 170.

[42] R. Adams, "Perceptions of innovations: exploring and developing innovation classification," PhD, School of Management Cranfield University, 2003.

[43] B. Mirkin, Clustering for Data Mining: A Data Recovery Approach (Computer Science and Data Analysis Series). Taylor & Francis Group, LLC, 2005, p. 266.

[44] A. C. Rencher, Methods of Multivariate Analysis, Second Edition ed. John Wiley & Sons, INC, 2002.