Author Verification of Nahj Al-Balagha

Raheem Sarwar and Emad Mohamed *

Research Group in Computational Linguistics, University of Wolverhampton, United Kingdom

December 14, 2023

Abstract

The primary purpose of this paper is author verification of the Nahj Al-Balagha, a book attributed to Imam Ali and over which Sunni and Shi'i Muslims are proposing different theories. Given the morphologically complex nature of Arabic, we test whether morphological segmentation applied to the book and works by the two authors suspected by Sunnis to have authored the texts, can be used for author verification of the Nahj Al-Balagha. Our findings indicate that morphological segmentation may lead to slightly better results than whole words and that regardless of the feature sets, the three sub-corpora cluster into three distinct groups using Principal Component Analysis, Hierarchical Clustering, Multi-dimensional Scaling and Bootstrap Consensus Trees. Supervised classification methods such as Naive Bayes, Support Vector Machines, k Nearest Neighbours, Random Forests, AdaBoost, Bagging, and Decision Trees confirm the same results, which is a clear indication that (a) the book is internally consistent and can thus be attributed to a single person, and (b) it was not authored by either of the suspected authors.

1 Introduction

Background. *Nahj Al-Balagha*, which can be translated into *The Clear Path to Eloquence* is a book in which the poet and author Al-Sharif Al-Radi (969-1015) collected the sayings, sermons, and letters of Ali ibn Abi Talib (601-661). The value of the book derives from two factors: (1) Ali is considered the most central figure person in the history of Shiite Islam and the first Imam. Ali was the cousin of the prophet Muhammad (PBUH) and his son-in-law. He was the fourth rightly guided caliph until his martyrdom, (2) the book is usually seen as the epitome of eloquence and is taught in schools. Many parts of the book have made it to popular culture.

*Corresponding Author

The book has always been a source of contention between Sunni Muslim scholars and Shiite Muslim scholars. There seems to be a consensus among Shiites that the content of Nahj Al-Balagha was authored by Ali. Sunnis, on the other hand, do not have such a consensus. The majority seem to dispute its authorship with some claiming that it was written by either Al-Sharif Al-Radi or his brother Sharif Al-Murtaza (966-1044) while some acknowledge that it is all authored by Ali. In his biography of Al-Murtada, Al-Qinnawaji writes:

> People have disagreed over the book of Nahj Al-Balagha, which contains the sayings of Imam Ali. Was it compiled by him or by his brother Al-Radi? It was also said that the sayings are not those of Imam Ali, but the one who compiled it and attributed it is the one who authored it. (Qinnawji, p.602)

The most comprehensive criticism of the attribution of Nahj AL-Balagha to Ali is from Sheikh Salih Al-Fawzan, a key Saudi cleric, who wrote a treatise rejecting the assumption that the book is Ali's based mainly on its lack of a chain of reporters and its containing passages that are contrary to what (Saudi) Sunni Muslim scholars believe to be the correct teachings of Islam.

Research Questions. The current study seeks to answer three questions:

1. Were the several passages in the book of Nahj Al-Balagh produced by the same person? and
2. Were the passages in the book of Nahj Al-Balgha authored by Ali, Al-Radi or Al-Murtaza?
3. Whether Arabic morphological segmentation helps with author verification?

Hussein made the first and only attempt to verify the authorship of Nahj Al-Balagha using function words [11]. However, it is not clear how many function words were used to carry out experiments and the study employs centroid hierarchical analysis, a single clustering method on its own might not be a reliable classification criterion.

Summary of Our Contributions. To answer the questions, we partition the book and the true writing samples of Al-Sharif Al-Radi and Al-Sharif Al-Murtaza into equal-sized chunks where the size of each chunk is 3000 tokens. We then extract stylometric features from each chunk. After the features extraction process, we conduct two main sets of experiments: (i) training machine learning models using stratified 5-fold cross-validation; and (ii) clustering analysis. Our experimental results show that the book was author by the same person (i.e., one single author) and neither Al-Sharif Al-Radi nor Al-Sharif Al-Murtaza is the true author of the book. It should be noted, however, that there is no way we can prove that the passages were authored by Ali or not as there is no other reference for what Ali said. All we can seek to prove is (a) whether the text is internally consistent and is thus written by the same or different authors, and (b) whether any of the two brothers wrote it or not.

The rest of the paper is organized as follows. Section 2 discusses the data (the corpus, prepossessing, the morphological complex nature of Arabic) and the methods (extraction of authorial features and clustering and classification approaches). Section 3 presents the results and explains the experimental findings while Section 4 contains the concluding remarks.

2 Data and Methods

To answer the questions above, we have collected data and employed author attribution methods as detailed below:

2.1 Data

This subsection describes the data collection and the morphologically complex nature of Arabic, the language of the corpus.

2.1.1 Data Collection

The data for this study come from three main sources:

1. The book of *Nahj Al-Balagha*, which is the main focus of this study. We use the version edited by Sheikh Faris Hassun. This edition is available online on a website dedicated to the book and its commentaries. The corpus has 342 passages varying in length with the shortest being 66 words and the longest 24525, with a standard deviation of 2645. The median is 580.

2. The book *Al-majazat al-nabawiyya* (Eng: Prophetic Metaphors) by Al-Sharif Al-Radi. This comprises passages totaling 15509 words. The book is a commentary on several Prophetic traditions.

3. For Al-Sharif Al-Murtaza, we use the book *The Epistles of Sharif Murtaza*, which has 74209 words (92618 with punctuation). The book expounds Sharif Murtaza's opinions on a number of legal and theological questions and has very few quotations.

2.1.2 A note on the Arabic language

Arabic is a morphologically rich language in which each word comprises a stem, zero or more prefixes and zero or more suffixes [6, 16]. Prefixes and suffixes may be inflections and may also be function words. For example, the word *fsnkfykhm* is made up of the conjunction f, the future particle s, the verb nkfy, and the singular second-person object pronoun k, and the third person plural masculine pronoun *hm*. The verb itself is made up of the suffix *n*, which denotes a plural first person subject and the stem *kfy*, and can be translated as *then we shall guard you against them* Due to this morphological complexity, Arabic has an artificially high type to token ratio, and it may be useful for several applications to reduce

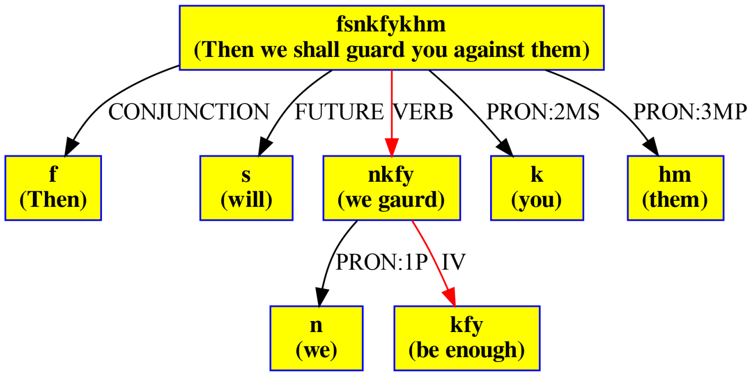


Figure 1: The structure of an Arabic word. The red arrows show the path to the verb stem, which is the lexical component. the prefix *n* is an inflection. All the other components are function words. 3MP = third person masculine plural; 2MS = second person masculine singular; 1P = first person plural

this ratio. It is also a necessary step if one is to obtain all the function words of Arabic since many of these are bound morphemes.

In order to solve the morphological complexity problem, we perform morphological segmentation, which delimits segment boundaries within words. The results of this process on the aforementioned word would be *f+s+n+ktb+h*, which enables us to treat each of these segments as a separate token. In the experiments below, we will see how effective this approach is in authorship attribution and verification. For the purposes of morphological segmentation, we use the Arabic-SOS package [17], which is specialized in Classical and pre-Modern Arabic. Arabic-SOS reports an accuracy of 99.5%, and we can confirm this very high accuracy on the corpus used in this article.

2.2 Methods

This section concerns the methods and experiments. We first discuss the nature of the Arabic language morphology the discuss pre-processing before we delve into the experiments.

2.2.1 Prepossessing

As with every task that involves language, there is a first step of prepossessing that needs to be done. While the morphological segmentation above can be considered prepossessing, it is only used in some experiments. The prepossessing we are more interested in involves the cleaning of the data to make it more amenable to analysis. This involves such things as separating or removing punctuation and external material. The most important of these is removing quotations. The books by Al-Radi and Al-Murtaza employ quotations is a means of supporting their opinions. They also explain and cite verses from the Qur'an and several Hadiths as well as sayings of other scholars of Islam. We have aimed at removing all these quotations through punctuational and linguistic cues. We note that no such process is one hundred percent correct, but we have identified several regular expressions for the process, and we believe that if any quotations remain, they are very few and should not bias the results. This step is necessary as we want to the linguistic content of these authors to be as unique to them as possible.

The other major cleaning process is the removal of footnotes. These footnotes were mostly by editors trying to explain unfamiliar words to the modern reader. To illustrate the importance of cleaning, the book Almajazat Al-Nabawiyya by Al-Sharif Al-Radi contains 97349 words. Removing the footnotes brought the number down to 55747, and removing quotes reduced it to 15509 words.

2.2.2 Extracting Authorial Features

In any authorship verification or attribution study, the researchers select the features based on which they run their supervised and unsupervised experiments. In this study we test three main sets of authorial features: (i) N most frequent words, (ii) N most frequent segments, (iii) N most frequent function words. We tested several values for N and 100 most frequent words, segments, and function words provided the best results. Therefore, for all the experimental studies illustrated in the following subsection, the value of N is fixed to 100.

- **N most frequent words.** The use of *N most frequent words* for authorship analysis is a well-tested approach [15]. However, they might be affected by the genre or theme of the text. As a result, we also test segments and function words based features to answer our research questions.

- **N most frequent segments.** This is very similar to N's most frequent words except that segment will be used instead of word. This is very similar to the top *N* words except that we use segments instead of words. For example, the word *fsnkfykhm* above will be treated as five different units: f, s, y, nkfy, k, and hm. This is better able to capture function words as many function words in Arabic are bound morphemes. This way, function words are better represented, which makes detecting author's style more straightforward.

- **N most frequent function words.** These features have also been used extensively in the literature. Since the seminal study by Mosteller and Waalace [18] on the disputed federalist papers, function words, not lexical ones, have been the focus of this investigative field. As we implied in discussing the word war, the words we want to use are non-contextual ones, words whose rate of use is nearly invariant under a change of topic. For this reason, the little filler words, called function words are especially attractive for discrimination purposes. The appeal of function words is summarised in [12] which lists the advantages as: (1) All authors use the same function words and are thus easily comparable, (2) there is always enough of them due to their high frequency, (3) they are less affected by the genre or theme of the text, and (4) they are not as under the conscious control of the author as are lexical words. Function words have also been used in various studies other than authorship attribution verification. The popular book *The Secret Life of Pronouns* documents such uses in emotion detection and analysis, lie detection and analysis, gender and age studies, and many other psychologically important research.

2.2.3 Experiments

In order to answer the questions posed in the introduction, the following experiments will be utilized:

- **Supervised Approaches**.

  We partition the book and the true writing samples of Al-Sharif Al-Radi and Al-Sharif Al-Murtaza into equal-sized chunks. We tried different values of the chunk size and 3000 tokens provided the best results, unless stated otherwise. We then extract stylometric features from each chunk. Specifically, we extract three sets of features from each chunk (i) 100 most frequent words, (ii) 100 most frequent segments (iii) 100 most frequent segments. Based on these three feature sets we conduct three different experiments to measure the effectiveness of these feature sets.

  In each experiment, we apply seven well-known classifiers to our dataset using stratified 5-fold cross-validation. These seven classifiers include Naive Bayes, k-Nearest Neighbours, Random Forest, AdaBoost, Bagging and Decision Trees which are described at the end of this subsection. Stratified 5-fold cross-validation provides train/test indices to split data into train/test sets. This cross-validation object is a variation of K-Fold that returns stratified folds. The folds are made by preserving the percentage of samples for each class. It generates test sets such that all contain the same distribution of classes, or are as close as possible, which is considered an ideal evaluation strategy for an unbalanced dataset. We used three well-known evaluation measures Precision, Recall, and F1 Measure.

  - **Naive Bayes.** For the Naive Bayes classifier, a set of training chunks is provided for each author. Each chunk is represented by a set of features. A new chunk is described by the same set of features, and the learner is asked to predict the author of the new chunk, assuming that the occurrences of the features are mutually independent [2].
  - **Support Vector Machines (SVM).** The SVM classifier is a supervised machine learning algorithm and has been extensively used in authorship attribution studies [9]. In the SVM algorithm, we plot each chunk as a point in 100-dimensional space with the value of each feature being the value of a particular coordinate. Then, we perform classification by finding the hyper-plane that differentiates the classes very well.
  - **K-Nearest Neighbors (KNN).** KNN is one of the popular classification methods used for the authorship attribution task. The KNN algorithms use chunks in the training data and classify chunks based on similarity measures (e.g. distance function). Classification is done by

a majority vote to its k neighbors. The chunk is assigned to the class which has the nearest neighbors.

– **Decision Trees, Random Forest, AdaBoost, and Bagging.** The goal of using a Decision Tree is to create a training model that can be used to predict the class by learning simple decision rules inferred training chunks. In a decision tree, each node represents an authorial feature, each link represents a decision(rule) and each leaf represents an outcome(author). For predicting an author for a chunk we start from the root of the tree. We compare the values of the root authorial print with the record's authorial print. On the basis of comparison, we follow the branch corresponding to that value and jump to the next node.

  A random forest algorithm consists of many decision trees. The 'forest' generated by the random forest algorithm is trained through bagging or bootstrap aggregating. Bagging is an ensemble meta-algorithm that can be helpful in improving the performance of the machine learning algorithms. The random forest establishes the outcome based on the predictions of the decision trees. It predicts by taking the average or mean of the output from various trees [4, 13].

  Bagging is a statistical estimation technique where a statistical quantity like a mean is estimated from multiple random chunks of data (with replacement). Multiple random chunks of training data are drawn with replacement and used to train multiple different machine learning models. Each model is then used to make a prediction and the results are averaged to give a more robust prediction [3].

  Adaboost is designed to use short decision tree models, each with a single decision point. Such short trees are often referred to as decision stumps. The first model is constructed as normal. Each instance in the training dataset is weighted and the weights are updated based on the overall accuracy of the model and whether an instance was classified correctly or not. Subsequent models are trained and added until a minimum accuracy is achieved or no further improvements are possible. Each model is weighted based on its skill and these weights are used when combining the predictions from all of the models on new data [8].

• **Unsupervised Approaches.**

  We apply four popular unsupervised approaches on our dataset including Principal Component Analysis, Multi-dimensional Scaling, Hierarchical clustering, and Bootstrap Consensus Tree.

  – **Hierarchical clustering (HC).** Hierarchical clustering algorithm groups stylistically similar chunks into groups called clusters. It produces a set of clusters at the end where each cluster is distinct from each other cluster, and the chunks within each cluster are broadly stylistically similar to each other. Hierarchical clustering starts by treating each chunk of text as a separate cluster. Then, it repeatedly executes the following two steps: (1) identify the two clusters that are closest together, and (2) merge the two most similar clusters. This iterative process continues until all the clusters are merged together. The main output of Hierarchical Clustering is a dendrogram, which shows the hierarchical relationship between the clusters. Eder's Delta is used as a distance measure between two chunks as it has proven to be effective for morphologically complex languages.

– **Principal Components Analysis (PCA).** Principle component analysis is a popular data visualization technique [10, 7]. The first step in this analysis is to quantify texts, i.e. convert words, segments or function words to numbers. This is done by putting together a most-frequent word, segment or function word list and by calculating the frequency with which each word appears in the texts. These values are then put into a matrix.

  In order to generate these graphs, this calculation is done with all the words, segments, or function words of all the chunks. As a result, what used to be words, segments, or function words are now transformed into matrices comprised of numbers. After generating the matrices, what follows is the generation of a matrix called the "covariance matrix" (or correlative) comprised of the variance present in the previous matrices. Variance can be explained as the distance between the values in the matrix and their average where Eder's Delta is used as a distance measure.

  After the covariance matrix is found, the eigenvectors are calculated. Eigenvectors are vectors found through mathematical equations in order to show the direction of the variance. Each eigenvector has an eigenvalue, which accounts for the magnitude of the variance. The two eigenvectors of the highest eigenvalue are then used as axes in which the data is projected in a two-dimensional space (two-dimensional since we only have two axes). In other words, the lower axis and the left axis you see are the eigenvectors that have been calculated. The percentage in parenthesis accounts for how much of the total variance is being projected. That way, the numeric values into which our chunks have been transformed are projected in a graph.

– **Multi-dimensional Scaling (MDS).** Multidimensional Scaling is a popular information visualization technique that allows us to visualize distances between a set of chunks[14]. Applying this technique to a matrix containing the distances (computed using Eder's Delta) between every chunk in the dataset will result in a set of two-dimensional points. The intuition is that texts that are similar in writing style, and thus probably written by the same author, are "close" to one another and "far" from texts written by another author

– **Bootstrap Consensus Trees (BST).** Several methods such as Hierarchical Cluster Analysis or Multidimensional Scaling can be used to find groups of stylistically similar works. There are several limitations of these methods including sensitivity to noise and outliers in the data and difficulty when handling with different sizes of clusters. Another limitation of these methods is that the order of the data has an impact on the final results. However, Bootstrap Consensus Network (BST) aims at overcoming its limitations with the help of network analysis [5]. Firstly, the algorithm establishes for every single node a strong connection to its nearest neighbor (stylistically similar text) and two weaker connections to the first and second runner-up. As a result, the final network will contain a number of links, some of them being close (most similar), some others revealing weaker connections between samples resulting in a consensus tree [19, 20, 1].

2.2.4 Authorship Verification and Validation

In this part, we compare Nahj Al-Balagha against the writings by the brothers Al-Sharif Al-Radi and Al-Sharif Al-Murtaza.

3 Results and Discussion

In this section, we present our findings based on the experiments on Nahj Al-Balagha. We then discuss the various aspects of these results.

3.1 Answer to question I

Question 1 raised the issue of whether the book *Nahj Al-Balagha* was authored by the same person. In order to answer this question, we use measures of internal consistency using both supervised and unsupervised learning.

As can be seen from experimental results given in Tables 1, 2, and 3 , the classification accuracy reaches 99%, which means that the classifiers are confident that these passages have the exact same (or very similar) features, and that they belong to the same author. We can also observe this very clearly in the unsupervised results shown in Figures 2, 4 and 3 where the Nahj Al-Balagha segments neatly cluster together, far from the other segments. We can thus conclude that the book of *Nahj Al-Balagha* was most likely authored by a single person, although we cannot yet claim that this author was Ali ibn Abi Talib given that there is no other Ali's gold standard to check it against. We can, however, test whether it was authored by either of the brothers Sunnis usually attribute the book to, which we handle next.

3.2 Answer to question II

Question II asked whether it was likely that any of the two brothers, Al-Radi and Al-Murtaza, authored the book they attributed to Ali. Both the classification and clustering results make it clear that this is not the case. The classification experiments can discriminate the three sources with an accuracy of 99%. It does actually seem that the choice of the classification algorithm may not be of any major consequence since all algorithm perform more or less the same with the exception of Decision Trees, which vary between 80% and 98%. Support Vector Machines and ensemble-learning algorithms. especially Random Forest, are perfect classifiers on this specific problem.

3.3 Answer to question III: the impact of morphological analysis on Arabic authorship attribution

Another algorithmic question is whether Arabic morphological segmentation helps with authorship attribution. Segmentation was required for the extraction of function words, so this specific experiment will be excluded. Apart from this, when we use the top 100 words or the top 100 segments, we find that using morphological segments performs better, or at least as well as, using whole words, which gives a case of support for performing morphological segmentation in general. The only exception to this observation is in the case of using Decision trees, which do not perform as well with morphological segments.

The use of function words (which are morphological segments) is better than extracting the top 100 segments regardless of their function/lexical status.

Classifier	Precision	Recall	F-Measure
Naive Bayes	0.987	0.984	0.986
SVM	0.991	0.996	0.999
K-NN	0.981	0.987	0.989
Random Forest	0.997	0.999	0.998
AdaBoost	0.999	0.992	0.995
Bagging	0.973	0.971	0.969
Decision Trees (J48)	0.972	0.972	0.972

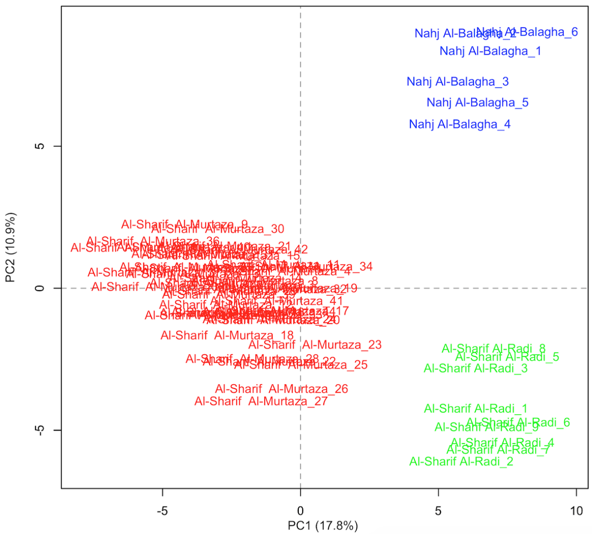
Table 1: Supervised Approaches using most frequent words as authorial features.

Classifier	Precision	Recall	F-Measure
Naive Bayes	0.999	0.998	0.999
SVM	0.991	0.994	0.992
K-NN	0.999	0.999	0.999
Random Forest	0.996	0.996	0.996
AdaBoost	0.993	0.993	0.993
Bagging	0.985	0.984	0.984
Decision Trees (J48)	0.807	0.800	0.802

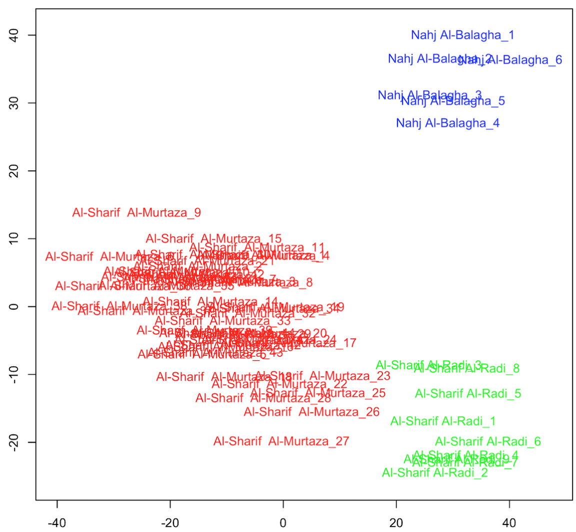
Table 2: Supervised Approaches using most frequent segments as authorial features.

Classifier	Precision	Recall	F-Measure
Naive Bayes	0.997	0.999	0.998
SVM	0.998	0.998	0.998
K-NN	0.999	0.999	0.999
Random Forests	0.993	0.999	0.997
AdaBoost	0.997	0.996	0.997
Bagging	0.989	0.989	0.989
Decision Trees (J48)	0.949	0.946	0.947

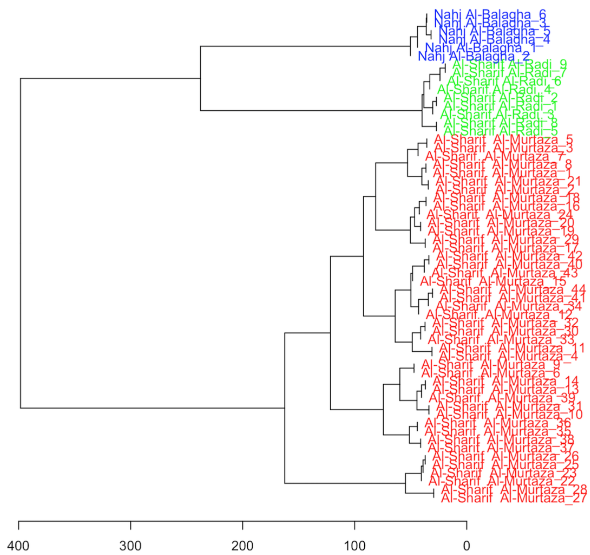
Table 3: Supervised Approaches using most frequent function words as authorial features and Stratified 5-fold Cross Validation where chunk size is 400 tokens to obtain reliable stylometric information from each chunk. Filtering out the content words results in a small number of chunks which leads us to reduce the chunk size for this experiment.



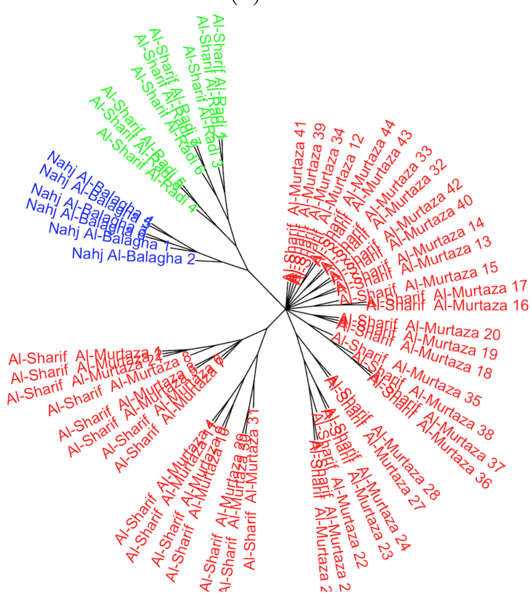
(a) PCA



(b) MDS

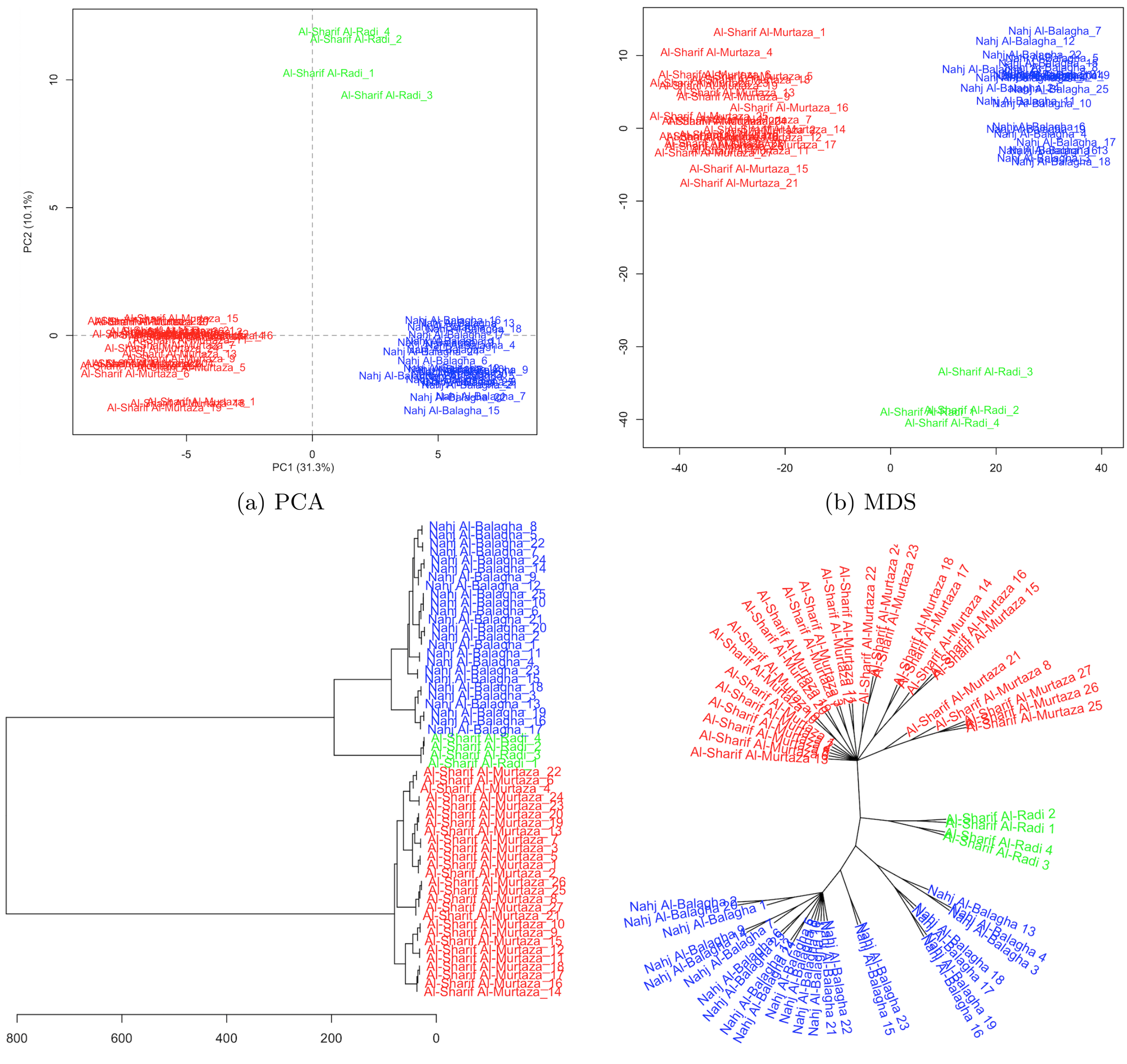


(c) HC



(d) BST

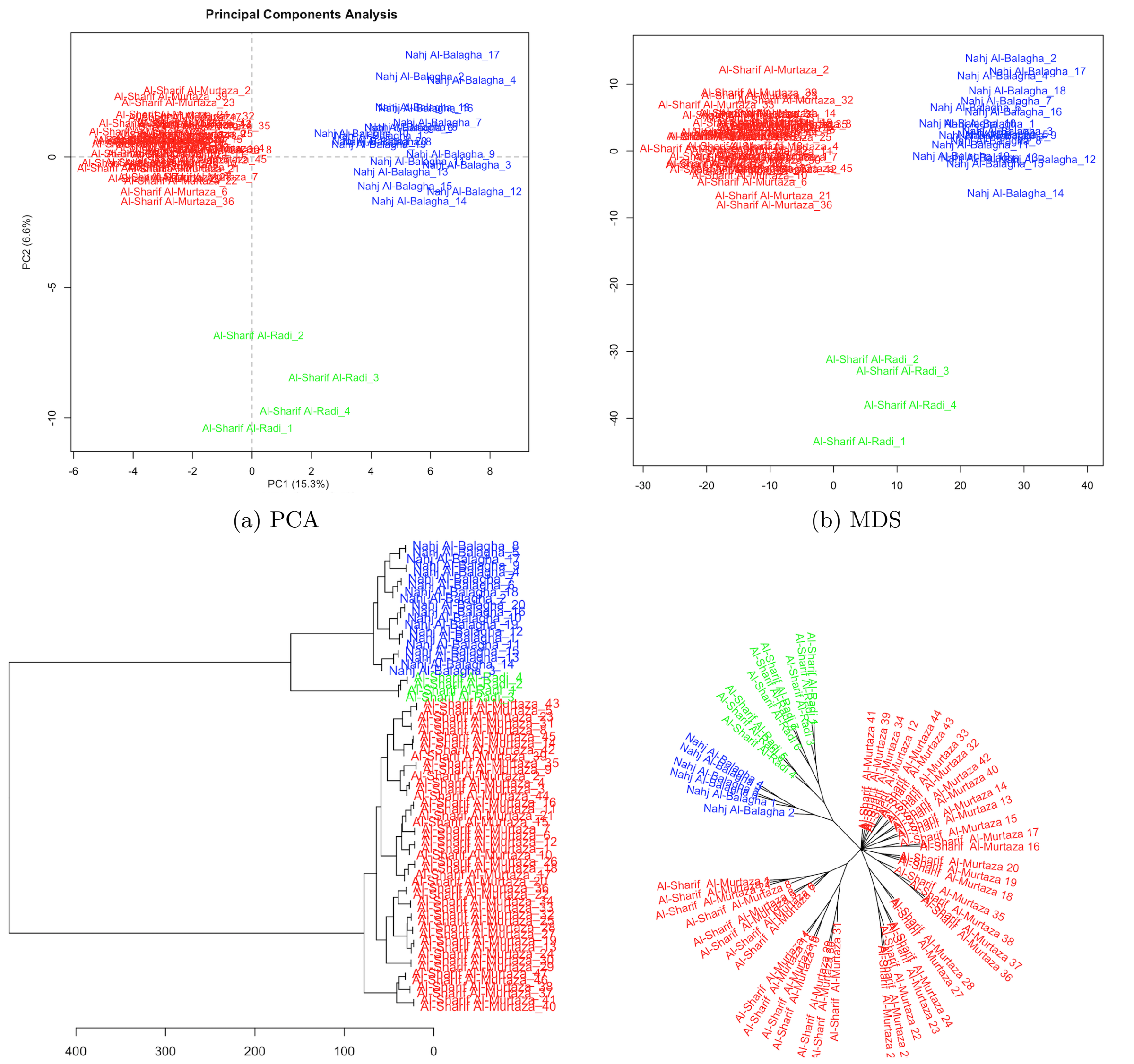
Figure 2: Unsupervised approaches using most frequent segments as authorial features.



(a) PCA (b) MDS

(c) HC (d) BST

Figure 3: Unsupervised approaches using most frequent words as authorial features.



(a) PCA

(b) MDS

(c) HC

(d) BST

Figure 4: Unsupervised approaches using most frequent function words as authorial features where the size of each chunk is 2K tokens.

3.4 Effect of the feature set size and the chunk Size

We also vary the chunk size and number of features to investigate their effect on the performance of the machine learning methods where segments are used as the authorial features. As can be seen from Figures 5 and 6 the 100 most frequent segments and 3000 tokens per chunk result in the best performance, respectively.

4 Conclusion

While Nahj Al-Balagha is a controversial book as Sunnis and Shi'is are vying over its authenticity, our computational analysis shows that: (1) The book is internally consistent and can thus be safely attributed

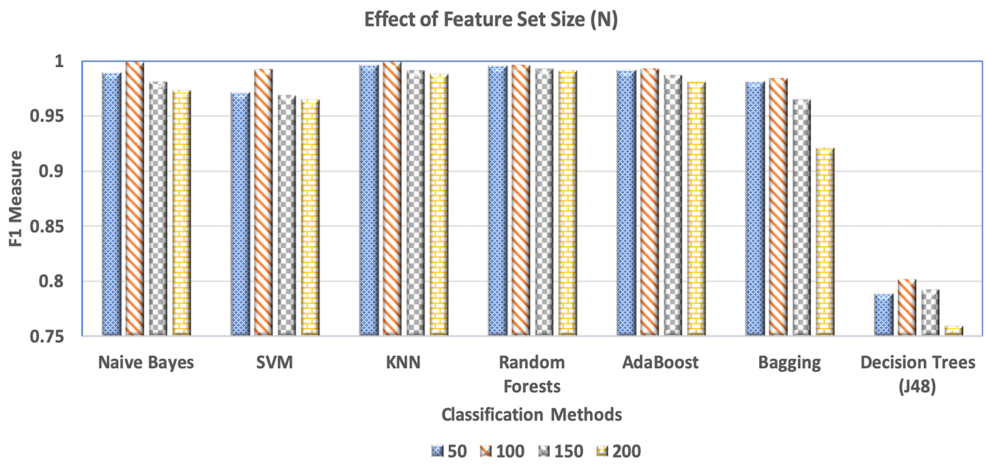


Figure 5: The Effect of Feature Set Size

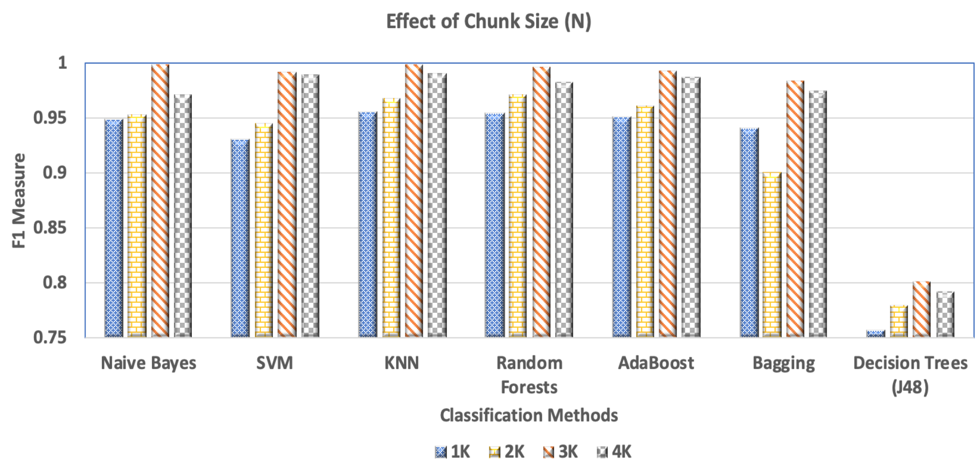


Figure 6: The Effect of Chunk Size

to one person, (2) it was not authored by Sharif Radhi, and (3) it was not authored by Sharif Murtaza. We have used multivariate analysis in different forms, two forms of data in Arabic (whole words and segmented), and several clustering and supervised machine learning algorithms, and in all cases, it is obvious that the data fall into three different clusters. The data can be so easily discriminated that we reach a classification f1 score with most algorithms and data representations.

While we have used function words/segments, most frequent words and segments, and standard algorithms for this specific task, in the future, we will devote a special study to the examinations of which feature set, and which algorithms work best for Arabic Authors Attribution in the presence of a gold standard data-set which we are currently creating.

References

[1] Sara Ali, Faisal Mehmood, Yasar Ayaz, Muhammad Sajid, Haleema Sadia, and Raheel Nawaz. An experimental trial: Multi-robot therapy for categorization of autism level using hidden markov

model. *Journal of Educational Computing Research*, page 07356331211040405, 2021.

[2] Alaa Saleh Altheneyan and Mohamed El Bachir Menai. Naïve bayes classifiers for authorship attribution of arabic texts. *Journal of King Saud University-Computer and Information Sciences*, 26(4):473–484, 2014.

[3] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[4] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[5] Maciej Eder. Computational stylistics and biblical translation: how reliable can a dendrogram be. *The translator and the computer*, pages 155–170, 2013.

[6] Abdelhamid Elewa. Authorship verification of disputed hadiths in sahih al-bukhari and muslim. *Digital Scholarship in the Humanities*, 34(2):261–276, 2019.

[7] Mel Evans and Alan Hogarth. Stylistic palimpsests: Computational stylistic perspectives on precursory authorship in aphra behn's drama. *Digital Scholarship in the Humanities*, 36(1):64–86, 2021.

[8] Yoav Freund, Robert E Schapire, et al. Experiments with a new boosting algorithm. In *icml*, volume 96, pages 148–156. Citeseer, 1996.

[9] Robert Gorman. Author identification of short texts using dependency treebanks without vocabulary. *Digit. Scholarsh. Humanit.*, 35(4):812–825, 2020.

[10] Hassina Hadjadj and Halim Sayoud. Arabic authorship attribution using synthetic minority oversampling technique and principal components analysis for imbalanced documents. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 15(4):1–17, 2021.

[11] Khalid Shakir Hussein. Authorship verification in arabic using function words: A controversial case study of imam ali's book peak of eloquence. *International Journal of Humanities and Arts Computing*, 13(1-2):223–248, 2019.

[12] Mike Kestemont. Function words in authorship attribution. from black magic to theory? In *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*, pages 59–66, Gothenburg, Sweden, April 2014. Association for Computational Linguistics.

[13] Peerat Limkonchotiwat, Wannaphong Phatthiyaphaibun, Raheem Sarwar, Ekapol Chuangsuwanich, and Sarana Nutanong. Domain adaptation of thai word segmentation models using stacked ensemble. Association for Computational Linguistics, 2020.

[14] Fernanda López-Escobedo, Julián Solorzano-Soto, and Gerardo Sierra Martínez. Analysis of intertextual distances using multidimensional scaling in the context of authorship attribution. *Journal of Quantitative Linguistics*, 23(2):154–176, 2016.

[15] Kim Luyckx and Walter Daelemans. Authorship attribution and verification with many authors and limited data. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 513–520, 2008.

[16] Emad Mohamed. Morphological segmentation and part of speech tagging for religious arabic. In Ali Farghaly and Farhad Oroumchian, editors, *Fourth Workshop on Computational Approaches to Arabic-Script-based Languages, CAAS@AMTA 2012, San Diego, California, USA, October 28 - November 1, 2012*, pages 65–71. Association for Machine Translation in the Americas, 2012.

[17] Emad Mohamed and Zeeshan Ali Sayyed. Arabic-sos: Segmentation, stemming, and orthography standardization for classical and pre-modern standard arabic. In *Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage*, DATeCH2019, page 27–32, New York, NY, USA, 2019. Association for Computing Machinery.

[18] Frederick Mosteller and David L. Wallace. Inference in an authorship problem. *Journal of the American Statistical Association*, 58(302):275–309, 1963.

[19] Jan Rybicki. The great mystery of the (almost) invisible translator. *Quantitative Methods in Corpus-Based Translation Studies: A practical guide to descriptive translation research*, 231, 2012.

[20] Jan Rybicki and Magda Heydel. The stylistics and stylometry of collaborative translation: Woolf's night and day in polish. *Literary and Linguistic Computing*, 28(4):708–717, 2013.