

Towards Automated *Fiqh* School Authorship Attribution

Maha Al-Yahya

College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia

Abstract. The word *Fiqh* (Islamic jurisprudence) refers to the body of Islamic law (Shari’ah). A large volume of *Fiqh* literature has been generated over the past thirteen hundred years, some of which texts have unknown authors. The importance of identifying the *Fiqh* School emanates from its importance in offering an authenticated interpretation of fundamental sources.

The traditional method for identifying the *Fiqh* School for a certain text is either by knowledge of the school affiliation the author or by close reading of the text by *Fiqh* scholars. This method is costly in terms of the time and human effort involved. An alternative to this manual approach is automated identification of *Fiqh* school texts using stylometric analysis. In this study we investigate the extent to which stylometric features can be used as predictors for *Fiqh* school authorship of a given text. We explore a corpus of Arabic *Fiqh* texts using unsupervised cluster analysis and supervised machine learning.

The results of our study show that the *Fiqh* schools have distinctive text style features that can be used to indicate authorship. The observations from the cluster analysis experiment using a number of different distance measures are visualized using network graphs. The best clustering in terms of *Fiqh* school division was achieved by the Classic Delta distance measure and Eder’s Delta distance measure. The results from the supervised experiment comparing the four classification algorithms: Support Vector Machines (SVM), Naïve Bayes (NB), K-Nearest Neighbor (KNN), and Delta show that supervised classification using SVM produces the highest average accuracy at 97.5% for the task of *Fiqh* school prediction.

Keywords: *Fiqh* school attribution, Stylometric Analysis, Arabic language, Cluster analysis, Supervised Classification

1 Introduction

The word *Fiqh* (Islamic jurisprudence) refers to the body of Islamic law (Shari’ah) that is derived from the two fundamental sources of Islam, the Qur’an (the divine word of God) and the Hadith (the authenticated prophetic teachings and practices). In the context of Islamic law (*Shari’ah*) studies, the importance of identifying the *Fiqh* School

emanates from the importance of Sharia law in the everyday lives of Muslims. There exists a large volume of Fiqh literature generated over thirteen hundred years, for some of which the author is unknown, and thus the Fiqh School is also unknown. To be able to utilize these texts in jurisprudence studies, it is vital to be able to recognize whether a given text is of the Fiqh School. The traditional method for identifying the Fiqh School for a certain text is either by knowledge of the text author school or by close reading of the text by Fiqh scholars. This method is costly in terms of the time and human effort involved. An alternative to this manual approach is to use automated identification of Fiqh school texts using stylometric analysis.

In this study we investigate the extent to which stylometric features can be used as a predictor for Fiqh school authorship of a given text. We explore a corpus of Arabic Fiqh texts from the four Sunni Fiqh schools (the Maliki, the Hanafi, the Shafi'i, and the Hanbali) using cluster analysis and supervised machine learning techniques. The stylometric features used are the most frequent words in the corpus (MFWs) and the cluster analysis is applied using a number of distance measures. For the supervised experiment, four machine learning algorithms are applied: Support Vector Machines (SVM), Naïve Bayes (NB), K-Nearest Neighbor (KNN), and Delta Classifier (specifically designed for the humanities literature [1]), are applied and compared.

The remainder of this article is organized as follows: section 2 presents background and related work in the area of stylometric analysis. Section 3 presents the methodology and experiment. Section 4 presents the results and evaluation and finally, section 5 presents conclusions and future work.

2 Related Work

Stylometric analysis is a form of computational modeling, more precisely it is a “quantitative approach to textual corpora” [2]. It involves “the extraction of the most appropriate features which can provide quantitative information about an author’s style” [3]. Stylometry, the analysis of authorship style [4], is based on the assumption that it is possible to identify an author’s style by studying the language the author uses – the stylometric features [5]. Stylometric analysis originated in the field of authorship attribution; however, it has been applied to other fields including textual forensics [6] [7], literary influence [8], genre detection [8], political affiliation [9], and comparative literature and translation studies [2].

In this study we leverage stylometric analysis to explore the Fiqh schools writing style in the field of Fiqh studies and investigate how quantitative methods can be applied in Islamic jurisprudence research. The problem of Fiqh school attribution can be viewed as an “author profiling” problem. Author profiling refers to a characteristic or property of the author, as Juola [10] points out “determining any of the properties of the author(s) of a sample of text”. Authorship profiling can be defined as “the analysis of a document to determine certain demographics such as gender without directly identifying the author” [4].

There are limited studies in the area of author profiling for Arabic. The work presented in [9] describes text classification for the task of political orientation determination. The authors study the problem of political orientation classification from both a text classification approach and a stylometric approach, and compare the results. The dataset they used includes articles and posts are collected from social networks and forums. For the stylometric approach, the authors select a set of features including lexical features (characters and words), syntactic features, structural features, and content-specific words. They conclude that the text classification approach is superior, with an accuracy of 83.9%, compared to the stylometric approach, with an accuracy of 60%. They suggest that different kinds of feature reductions, such as stemming and feature selection, have negative effects on the results.

Another study on ideological and organizational affiliation classification for Arabic text is presented in [11]. The authors use a stylometric approach to address the problem. The dataset is composed of 2 corpora, one for organizational affiliation with documents categorized into 4 categories, and one for ideological streams categorized into four different categories, both manually labeled. The features selected were the set of 1000 most frequent words (MFWs) of the corpus. No stemming has been applied. The study employed the Bayesian multi-class regression (BMR) supervised learning method. For organization identification, the accuracy is 100% with 1000 MFW, and with 82 MFW the accuracy is 80%. The results indicate that stylometric features represent a robust and reliable method for categorization of Arabic text into organizational or ideological streams. However, the authors did not apply the classification to other classification models.

The study presented in [12] describes an experiment on author profiling for Arabic emails. The study presents machine learning classifiers for different author traits including age, gender, education, and psychometric traits. Comparing the classifiers, the best performance is achieved by SVM and Bagging. Another study which focuses on author gender identification for Arabic is presented in [13]. The authors view the problem as a classification problem and a number of classification algorithms are applied on an Arabic dataset using the bag-of-words approach for feature extraction. The results show high accuracy for the SVM classifier in regard to the gender identification task.

3 Method

The process of stylometric analysis for Fiqh attribution is shown in Figure 1. Using the Fiqh dataset, the text is first pre-processed. Pre-processing involves tasks such as tokenization, stemming, part of speech (POS), removing non-alphabetic characters and spaces, and converting uppercase letters to lowercase. For our study, the only pre-processing applied is tokenization and removing of diacritics as this study is intended to be a baseline for future studies in which other forms of pre-processing can be applied to the corpus to determine the impact of such on performance.

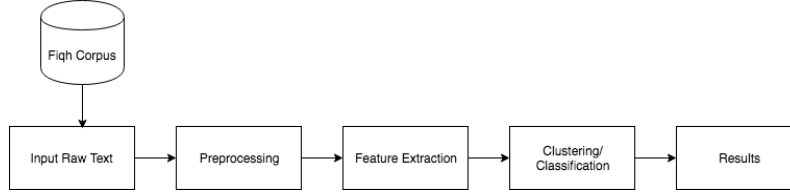


Fig. 1. Stylometric Analysis Process

The next task of the stylometric analysis is feature extraction. Features are extracted from the pre-processed texts. For the experiment in this study, function words are used as features, which are the MFWs in the dataset, which have been reported as a baseline in the past and shown to yield good results in stylometric studies [10]. After feature selection, the analysis is performed and finally the results are presented. In the case of clustering, a clustering algorithm is selected and the clustering is applied on the complete dataset. The results are presented in a diagram, and in our study we represent the results as a network graph. In the case of classification, the corpus is divided into two sets; a training set to train the classifier and a test set for testing the accuracy of the Fiqh school prediction. The results of the classification are represented in the form of the accuracy of prediction.

The package Stylo [1], developed for the R environment for statistical computing [14], is used for the pre-processing, the extraction of the features, and the application of the clustering and classification analyses to the Fiqh dataset.

3.1 Dataset

The dataset for the study was extracted from *Almaktabah Al Shamela* website [15], which is an online digital library of Arabic texts from the early pre-Islamic era to the modern period covering a variety of subjects including language, religion, and science. A Fiqh corpus was developed by extracting all Fiqh texts available in the library¹, organized into the four major Sunni Fiqh schools -Hanafi, Maliki, Shafi'i, and Hanbali. The documents were converted into raw text format (txt) with no diacritics. No form of pre-processing or stemming was performed on the dataset except tokenization. The dataset includes a total of 135 books from the four Sunni Fiqh schools Hanbali, Hanafi, Shafi'i, and Maliki, the representation of each school in the corpus is 29%, 26%, 26%, and 19% respectively.

3.2 Cluster Analysis

Clustering is the process of dividing data into meaningful groups (clusters). In clustering analysis, a measure of nearness is computed between documents [16]. The choice

¹ As of November 2017.

of distance measure is an important step in clustering as it defines how the similarity between the two texts is computed and it will influence the shape of the clusters. There are a number of distance measures used in clustering for stylometric analysis which include: (1) Classic Delta (originally the Burrows' delta), (2) Eder's Delta, (3) Eder's Simple Delta, (4) Argamon's Delta, (5) Canberra distance, (6) Manhattan distance, (7) Euclidean distance, and the (8) Cosine distance. The cluster analysis method used is the bootstrap consensus tree (BCT) approach which is implemented in the Stylo package using 100-1000 MFWs as features. In this study cluster analysis is applied using the eight distance measures, and the results are compared to identify the best clustering among the four Fiqh schools.

Comparing the resulting network graphs, the best clustering in terms of Fiqh school division was achieved by the Classic Delta distance measure and Eder's Delta distance measure, as shown in Figure 2, where the color denotes the actual Fiqh School (green for Hanbali, orange for Hanafi, purple for Shafi'i, and blue for Maliki). From the graph we can see that the texts are generally grouped into four clusters. Two of the clusters show a clear grouping of the Hanafi (orange) and Shafi'i (purple) Fiqh schools. For the other two clusters, one contains mixed texts from all four schools (central cluster), and the other shows two sub-clusters of Hanbali (green) and Maliki (blue) texts intermixed in one large cluster. The observations resulting from this initial cluster analysis experiment show that the Fiqh schools have distinctive style features that can be indicative of Fiqh school affiliation of the author.

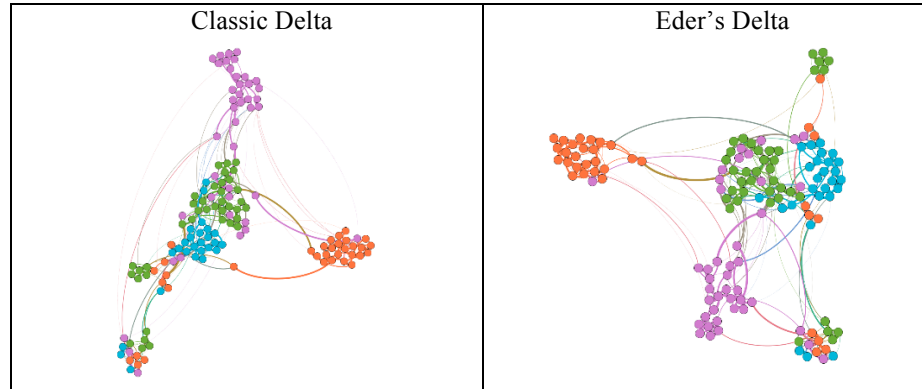


Fig. 2. Fiqh School Clustering using Classic Delta and Eder's Delta

3.3 Classification

For classification, the dataset is divided into a training set (80%) and a test set (20%). Four classification algorithms are applied on the training set, and the resulting classifier is used to predict Fiqh school authorship in the test set. The predication accuracy is measured as the percentage of correct Fiqh school predictions. Stylo has a number of

implementations for supervised machine learning classification algorithms which include Support Vector Machines (SVM), Naïve Bayes (NB), K-Nearest Neighbor (KNN), and Delta. A classifier was trained on the training dataset using each of these algorithms. For reliability, and in order to obtain more generalizable results, a 20-fold cross-validation was performed using the 100-1000 MFWs as features for the classifiers, with increments of 100 MFWs. The average accuracy of all classifiers is computed. Table 1 shows the results.

Table 1. Fiqh School Classification Performance

	SVM	kNN, k=1	kNN, k=3	kNN, k=5	kNN, k=7	kNN, k=9	NB	Delta
Average accuracy	97.4%	81.1%,	76.3 %	61.9 %	63.3 %	65.2 %	81.1%	93.7%

4 Evaluation

Results of the cluster analysis experiment show that a text which belongs to a certain Fiqh school has style features that can be discriminative of its Fiqh school authorship. The visualization of the network graphs generated from the clustering experiment show that the Classica Delta and Eder's Delta distance measures were able to generate the best group divisions based on text style. Eder's Delta has been designed to work with highly inflected languages such as Arabic [1]. Moreover, the intermixed cluster showing the closeness of the Hanbali and Maliki texts might be due to the assumption that the methodology used for reasoning in the interpretation of Islamic resources is similar in these two schools.

Regarding the classification experiment, results show that the best accuracy was achieved by the SVM classifier (97.4%) when using the set of 100-1000 words as features. The Delta classifier achieved very good results, giving an average accuracy of 93.7%. The best results for the kNN is achieved for k=1 (81%). SVM is known to perform well and produces high accuracy for high-dimensional problems such as text classification [17]. The Delta classifier produced good results for the task of Fiqh school prediction, which might be due to the fact that this classifier has been tailored for use in humanities literature, of which the Fiqh corpus can be considered part.

5 Conclusion and Future work

The work presented in this paper is a contribution to the quantitative methods that can be deployed in Islamic jurisprudence research. The aim of the work is to evaluate the existence of stylistic features of Fiqh texts that can support the task of Fiqh school attribution. Two experiments have been performed on a dataset of 135 Fiqh texts belonging to four different Fiqh schools. The clustering experiment showed that the style of a text is indicative of Fiqh School, and the best clustering is achieved using the Classical Delta and Eder's Delta. For the classification experiment, the best accuracy of prediction of the classifier was achieved by the SVM classifier, giving a prediction accuracy of 97.4%. The results also indicate that the use of the 100-1000 MFWs as features is suitable for the task of Fiqh school prediction.

One of the aims of this work is to develop a baseline for further development on Fiqh School attribution studies. For example, the only pre-processing implemented in this study was tokenization and diacritics removal; however, other pre-processing tasks can be performed which include stemming, tagging with part of speech, morphological analysis, and named entity extraction. The effect of pre-processing on classification performance can be analyzed. Moreover, the dataset can be extended to include more Fiqh texts.

References

1. Mike, K., Rybicki, J., Eder, M.: Stylometry with R: a Package for Computational Text Analysis. *The R Journal*. 8, 107–121 (2016).
2. Wrisley, D.J.: Modeling the Transmission of al-Mubashshir Ibn Fātik's Mukhtār al-Ḥikam in Medieval Europe: Some Initial Data-Driven Explorations. *The Journal of Religion, Media and Digital Culture*. 5, 228–257 (2016).
3. Stamatatos, E.: A survey of modern authorship attribution methods. *J. Am. Soc. Inf. Sci.* 60, 538–556 (2009).
4. Neal, T., Sundararajan, K., Fatima, A., Yan, Y., Xiang, Y., Woodard, D.: Surveying Stylometry Techniques and Applications. *ACM Comput. Surv.* 50, 86:1–86:36 (2017).
5. López-Escobedo, F., Solorzano-Soto, J., Martínez, G.S.: Analysis of Intertextual Distances Using Multidimensional Scaling in the Context of Authorship Attribution. *Journal of Quantitative Linguistics*. 23, 154–176 (2016).
6. Rocha, A., Scheirer, W.J., Forstall, C.W., Cavalcante, T., Theophilo, A., Shen, B., Carvalho, A.R.B., Stamatatos, E.: Authorship Attribution for Social Media Forensics. *IEEE Transactions on Information Forensics and Security*. 12, 5–33 (2017).
7. Afroz, S., Brennan, M., Greenstadt, R.: Detecting Hoaxes, Frauds, and Deception in Writing Style Online. Presented at the May (2012).
8. Jockers, M.L.: *Macroanalysis: Digital Methods and Literary History*. University of Illinois Press, Urbana (2013).

9. Abooraig, R., Alwajeih, A., Al-Ayyoub, M., Hmeidi, I.: On the Automatic Categorization of Arabic Articles Based on Their Political Orientation. Presented at the September 22 (2014).
10. Juola, P.: Authorship Attribution. *Found. Trends Inf. Retr.* 1, 233–334 (2006).
11. Koppel, M., Akiva, N., Alshech, E., Bar, K.: Automatically Classifying Documents by Ideological and Organizational Affiliation. In: *Proceedings of the 2009 IEEE International Conference on Intelligence and Security Informatics*. pp. 176–178. IEEE Press, Piscataway, NJ, USA (2009).
12. Estival, D., Gaustad, T., Hutchinson, B., Pham, S.B., Radford, W.: TAT: an author profiling tool with application to Arabic emails. In: *Proceedings of the Australasian Language Technology Workshop 2007*. , Melbourne, Australia (2007).
13. Alsmearat, K., Al-Ayyoub, M., Al-Shalabi, R.: An extensive study of the Bag-of-Words approach for gender identification of Arabic articles. Presented at the November (2014).
14. R Core Team: *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2016).
15. Almaktabah Alshamela, available online <http://shamela.ws/> , (accessed January 2018) .
16. *Data Analysis and Data Mining: An Introduction*. Oxford University Press, Oxford, New York (2012).
17. Joachims, T.: Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In: *Proceedings of the 10th European Conference on Machine Learning*. pp. 137–142. Springer-Verlag, Berlin, Heidelberg (1998).