Computer Stylometric Comparison of Writings by Qassim Amin and Mohammed Abdu on Women's Rights

Ahmed Ibrahim Ahmed Omer and Michael P. Oakes
Research Institute of Language and Information Processing
University of Wolverhampton
Wolverhampton, England
a.omer@wlv.ac.uk

Abstract

Computer stylometry is the computer analysis of writing style. We use the computer stylometric techniques of Hierarchical Cluster Analysis, Principal Component Analysis and Machine Learning to examine the authorship of "The Liberation of Women" which is normally attributed to Qassim Amin. In particular we examine the assertion by Mohamed Emara that certain chapters of this book were written secretly by Mohammad Abdu, who was the Grand Mufti of Egypt. In our experiments, we consistently find that Qassim Amin is the more likely author of the disputed text. The experiments described in this paper were done using the "Stylometry in R" package of Eder et al. (2016).

1 Introduction

Women's liberation was started in Egypt in the nineteenth century when Egypt was ruled by Mohammed Ali (1769-1849), and at that time the first school for training women was established. The school was used to train women to become medical assistants. Later Mohammed Ali opened the first primary school for girls. Ali sent many students to France to study there and to become leaders in different positions in the government. One of the students who was sent to France to do a Ph.D. degree was Qassim Amin. Qassim Amin finished his degree in law and traveled to France to stay there for about four years. Amin's traditional view of society was altered, and he started to see Egyptian women's lives through different eyes. He started to believe that life in the Egyptian community would not be improved unless the status of the women in society could improve. He believed that the main reason for their inferior position in the Egyptian community was ignorance about women and the lack of education. Accordingly, in 1899 he introduced his book "The Liberation of Women" (Amin, 2000) and used both rational Islamic arguments and emotional arguments to put forward his view. In his book, Amin called for women's education, removing the veil, and reformation of marriage and divorce laws. Mohamed Emara who was born in December 1931 is an Islamic scholar, author, investigator, and member of the Islamic Research Academy at Al-Azhar, Cairo. He did his Master's degree in Islamic philosophy at the Faculty of Science at Cairo University. He also got his Ph.D. from the same University in 1975. Mohamed Emara published many books about the life of many Islamic scholars including Jamal al-Deen al-Afghani, Abdul Razzaq Sanhouri Pasha, Sheikh Mohammed al-Ghazali, Rashid Rida and Muhammad Abdu. Emara stated in his book "Islam and Women in Mohammed Abdu's opinion" (Emara, 1997) that many chapters of Qassim Amin's book "Liberation of Women" were written secretly by Mohammed Abdu (1849-1905) who was a teacher of Qassim Amin and a religious scholar, jurist, and liberal reformer. In this paper we used techniques of computer stylometry to examine whether it is more likely

that Mohammed Abdu wrote these chapters secretly or they were indeed written by Qassim Amin as generally thought.

2 Related Work

Discriminating between different authors is a challenging task, especially when there is a dispute about a text, and two or more authors have claimed that they have written this disputed text. Many studies in this field have tried to analyse the texts to find proof of authorship, and different approaches are used by researchers in the domain to find this evidence. A typical approach is finding the frequency of specific patterns appearing in the author's texts. These can be, for example, the frequency of words, the word lengths, or the sentence lengths. Shaker and Corne (2010) used a set of 104 function words to analyze texts derived from samples extracted from the website of the Arab Writers' Union (www.amu-dam.net). Their work was mainly inspired by Mosteller and Wallace's (1964) typology of English function words. By shifting their focus to function words, they were able to effectively capture the author's writing style regardless of the text's topic, as the use of these words in a text is normally unrelated to the topic, and yet there appear to be significant differences in the way different authors use them in writing. Ouamour and Sayoud (2012) tested different character and word features using an SMO-SVM classifier on text samples extracted from textbooks. These features included character bigrams, character trigrams, character tetragrams, single words, word bigrams, word trigrams, word tetra grams, and rare words. The texts were collected from ten different authors who wrote their texts in the domain of travel. Kumar and Chaurasia (2012) used character n-grams, especially bigrams and trigrams, to solve the problem of authorship verification. The authors investigated the bigram in different positions in words. They tested bigrams in the initial, the middle position, and at the end of words. The tests were conducted for both English and Arabic texts. The results showed that the initial bigrams and trigrams were the most useful in accomplishing the task. Howedi and Mohd (2014) examined many features to classify authors according to style. The authors used different linguistic features such as character bigrams, character trigrams, single words, word bigrams, and rare words. The dataset used in these experiments was the same data as used by Siham and Sayoud (2012), the AAAT Corpus of ten ancient Arabic travellers. Al-Zubaidi and Ehsan (2017) used the 300 most frequent features to predict the authors in Arabic texts. The dataset consisted of 18 books written by three old Arabic philosophers, Ibnjuzia, Sakhawy, and Tusi. Each one of the authors was represented by six books. Five books were used for training, and the sixth book was used for testing purposes. In this paper we used different linguistic features to discriminate between authors. To predict the authors of the texts, we represented each text sample by a vector. The vectors were numerical representations containing the frequency of the linguistic feature used in each sample. We then created a matrix where the columns corresponded to linguistic features (i.e. word frequencies), and the rows corresponded to individual texts. Finally, the classic Delta distance was used to calculate the distance between vectors. A matrix was produced of the distances between each pair of texts, and these distances were used as the basis of the clustering techniques we used, Hierarchical Agglomerative Cluster Analysis (HACA) and Principal Component Analysis (PCA). The initial vectors were also used as inputs to a set of machine learning techniques.

3 Corpus Description

To check whether the chapters included in the book "Liberation of Women" were written by Qassim Amin or Mohammed Abdu, we built a corpus containing their writings about women. The corpus contained three different sources. Two of these were written by Qassim Amin, and the last one was written by Mohammed Abdu. The first texts used in the experiments were extracted from the first book "The Liberation of Women" which were the chapters about women, and these were the chapters which Mohammed Emara stated were written by Mohmmed Abdu and included in Amin's book without any mention that they were Abdu's contribution to this book. Emara assumed that Amin was not qualified enough to discuss this topic and supported his view by citing new interpretations for verses from the Holy Quran in these chapters. He assumed that only Amin's teacher Mohammed Abdu was able to do this at that time. He also stated that the writing style of these chapters was more similar to Abdu's style than Amin's style. The second set of

texts used in the experiments, written by Amin, were extracted from the book "The New Women". This book, written later by Qassim Amin after the death of Abdu also spoke about women's rights and asked the community to give women more rights. The last texts contained in the corpus were a collection of Mohammed Abdu's texts which were written about the rights of women. The texts were extracted from a book written by Mohammed Emara to discuss Abdu's opinions on women's rights in Islam. The following table describes the texts used to build the corpus:

Book	Author	The extracted texts
The Liberation of Women	Qassim Amin	G_L1 to GL5
The New Women	Qassim Amin	G_N2 to GN5
Mohammed Abdu's Opinion on Women	Mohammed Emara	M_MAB2_1 to M_MAB2_5

Table 1: Corpus Description.

4 Experiments

In this case study we extracted various samples from each book to compare the style of the writing. Many linguistic features of the texts were then examined including the most frequent words, character 3 grams, and character 6 grams, to discriminate between Abdu and Qassim Amin. Two different methods for analysis were used to investigate the case which were Agglomerative Clustering, and Principal Component Analysis (Eder *et al.,* 2016). Machine learning techniques were also used to label the texts according to author.

4.1 Most Frequent Words Cluster

In this experiment we used the most frequent words as a feature set to discriminate between the two authors. This set of most frequent words is usually a set of the function words. This set can be used by the same author to write on two different topics because these words are topic independent. The most frequent hundred words were used to automatically cluster the texts according to style. The following graph shows that the texts which were extracted from the book "The Liberation of Women" was clustered beside the texts which were extracted from the book "The New Women" which confirm that these texts were written by Qassim Amin, and the texts written by Abdu were clustered in a different branch on the left as shown below:

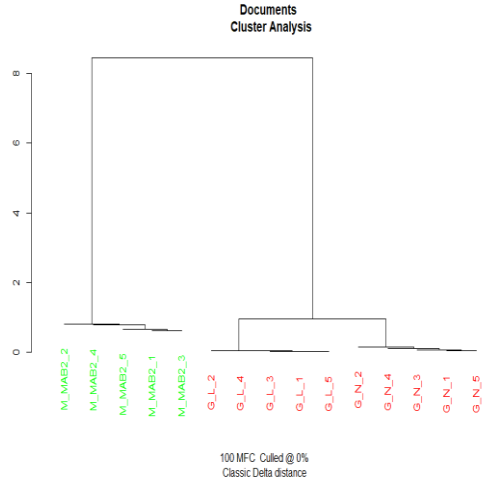

Figure 1: The most frequent words cluster

4.2 Character n-grams Clusters

In this experiment, we used the features character 3-grams and character 6-grams. We uploaded the corpus after doing the tokenization process to produce the corpus in the required format. The tokenizer made a window of a specific length and then cut the words into the required length assuming that the space between words is a character so that we can find a token consisting of one character from a word and another character from the next word, concatenated with a space. For example, the token [y m] can exist in the corpus among the character 3-grams tokens as a result of the two words (happy man). The following graphs show the results of the Agglomerative Clustering method using the features character 3-grams and character 6-grams.

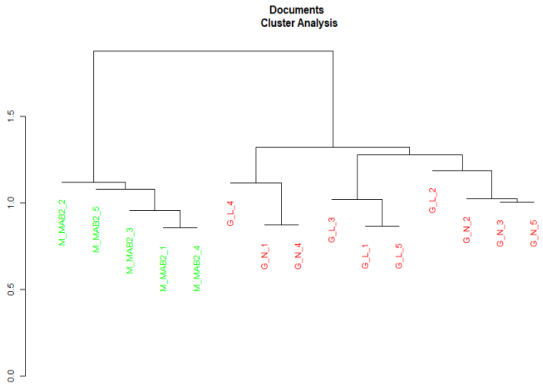

Figure 2: Character 3-grams cluster

From Figure 2 we can see that the texts from the two books which were written by Qassim Amin were clustered together while the texts from Mohammed Abdu were clustered in the leftmost branch of the tree. In addition, the samples from the two books written by Amin were mixed inside the cluster, and this shows that this feature was a very useful feature to capture the fingerprint of Qassim Amin. The same scenario occurred when we used character 6- grams as a feature set to discriminate between the two authors. Figure 3 below was produced by using character 6-grams as the feature set.

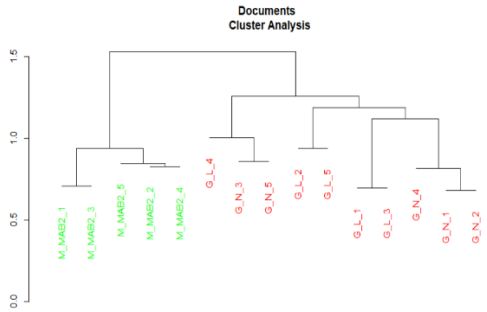

Figure 3: Character 6-grams cluster

4.3 Principal Components Analysis (PCA)

To confirm the results obtained by using the clustering technique using different feature sets, we used PCA analysis to find whether any groups of tokens occurred together in a specific group of documents more than the others so these could be used as a feature set to discriminate between the two authors. We ran the experiment using the 100 most frequent words as the feature set and found out that the first principal component was useful to sperate the two authors in a clear graph showing that the texts which were extracted from the book "The Liberation of Women" were more similar to the texts which were extracted from the book "The New Women" than the text by Muhammad Abdu. Figure 4 shows that the texts which were written by Amin are separated on the left-hand side and the features used to build the model were also shown in the graph. The most frequent words show that Mohammed Abdu uses the term "Man" (الرجل) while Qassim Amin use the term "Men" (الرجال) when they speak about the male gender.

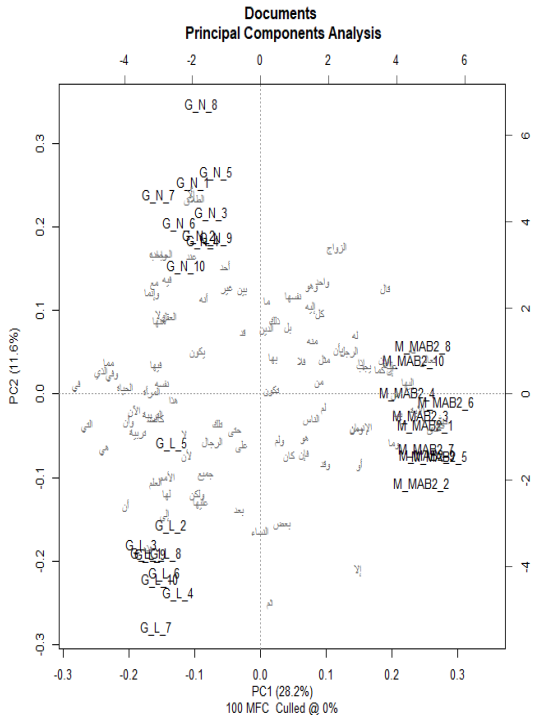

Figure 4: PCA using the most frequent words

In another experiment, we extracted the different morphemes which were available in the corpus by using the Farasa tools (http://alt.qcri.org/farasa/) to do word segmentation and the Stylo package (Eder et al., 2016) to find the most frequent features and used them as a feature set to discriminate between the two authors. The PCA showed that the different morphemes could be used to discriminate between the two authors as the texts of Qassim Amin were separated from Abdu's texts by the first principal component. The following list of morphemes (/ف / ب/ت/ و/ ل /هن /ال) were observed on Abdu's side while the following list of morphemes (ات/ نا/ها / ي/ة) were observed on Amin's side on the graph.

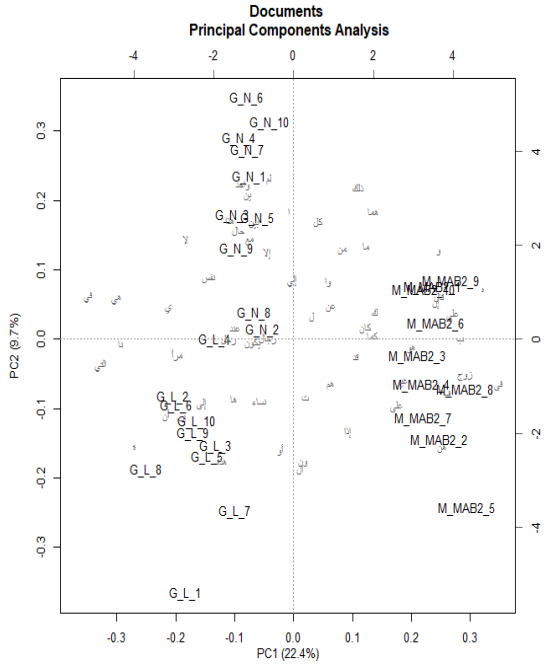

Figure 5: PCA using morphemes and Function words

4.4 Machine Learning Techniques

To check the results obtained by Agglomerative Clustering and PCA, we used machine learning techniques to find whither the chapters which were extracted from the book "Liberation of Women" would be classified or labelled as texts written by Qassim Amin or Mohammed Abdu. Ten samples were extracted from the book "The Liberation of Women" to be labelled by the classifier, and the texts from the book "The New Women", besides the texts by Mohammed Abdu, were used as a training corpus to extract the useful patterns which could be used to predict the disputed texts. Five different machine learning classifiers were then used to predict the author of the ten samples, and the results confirmed that the sample texts were written by Qassim Amin. The following table shows the classifiers used together with the accuracy achieved using the different classifiers:

Classifier	**Words**	**Character 6-gram**
NSC	90%	90%
SVM	90%	90%
Naïve Bayes	80%	70%
Delta	90%	80%
KNN	80%	70%

Table 2: Classifiers Accuracies

This table shows the different accuracies obtained by using the different classifiers. For example, the accuracy 90% means that 9 samples out of the ten were classified as written by Qassim Amin and one sampled was classified as written by Abdu.

5 Conclusion

In this paper, we introduced a case study about Qassim Amin's book "The Liberation of Women." Qassim Amin was a liberal reformer who advocated giving women more rights in the Arab and Muslim communities. He assumed that the lack of education for women could affect not only women but also the whole society. Later some scholars including Mohammed Emara (Emara, 1997) reported that there are some chapters included in Amin's book "The liberation of Women" which were written secretly by his teacher Mohammed Abdu. We decided to use the Computational Stylometry to investigate this argument and to find whither Abdu wrote these chapters or not. To do the experiment we built a corpus which contained different texts extracted from Amin and Abdu's books. The disputed texts together with other texts extracted from the book "The New Women" were used to compare Amin's style against the texts which were extracted from Abdu books on the same topic. Different features were used to investigate the texts' writing style including the most frequent words, Character 3-grams, and Character 6-grams. The character n-grams were very useful features as they captured the fingerprint of the author. The extracted texts from the two books written by Amin clustered together in one branch in the clustering tree. To confirm the results obtained using the clustering technique we used Principal Component Analysis (PCA) to analyse the texts. The results confirmed that the disputed texts were written by Amin and not by Mohammed Abdu, and the texts from Amin were perfectly separated by the first principal component. We then used machine learning techniques to check whether we can successfully label the texts according to the author's writing styles. To do that we extracted samples from "Liberation of Women" and used them as a testing corpus to be labeled by the classifiers. We also extracted samples from "New Women" together with texts written by Mohammed Abdu on women's rights to form a

training corpus. Five different classifiers were then used to find the fingerprint of the authors from the training corpus to build the model and to automatically predict the author of the texts available in the testing corpus. The results showed that the disputed texts were more similar to Qassim Amin's style than Abdu's style. In the future we would like to extend our experiments by adding more works written by Abdu and Qassim, possibly on different topics, as the most frequent words set of features is topic independent.

References

Abdelali, Ahmed, Kareem Darwish, Nadir Durrani, and Hamdy Mubarak. "Farasa: A fast and furious segmenter for arabic." In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations, pp. 11-16. 2016.

Mustafa, Tareef Kamil, Ammar Adil Abdul Razzaq, and Ehsan Ali Al-Zubaidi. "Authorship Arabic Text Detection According to Style of Writing by using (SABA) Method." Asian Journal of Applied Sciences (ISSN: 2321–0893) 5, no. 02 (2017).

Amin, Qasim. "the Liberation of Women." In Modernist and Fundamentalist Debates in Islam, pp. 163-181. Palgrave Macmillan, New York, 2000.

Amin, Qasim, Qāsim Amīn, and قاسم أمين. The liberation of women: And, the new woman: Two documents in the history of Egyptian feminism. American Univ in Cairo Press, 2000.

Eder, Maciej, Jan Rybicki, and Mike Kestemont. "Stylometry with R: a package for computational text Analysis." R journal 8, no. 1 (2016).

Emara, Mohammed., 2004. Islam and Women in Mohammed Abdu's Opinion. Dar al-rashad, Cairo.

Hadjadj, Hassina, and Halim Sayoud. "Towards an authorship analysis of two religious documents." In 2016 8th International Conference on Modelling, Identification and Control (ICMIC), pp. 369-373. IEEE, 2016.

Howedi, Fatma, and Masnizah Mohd. "Text classification for authorship attribution using Naive Bayes classifier with limited training data." Computer Engineering and Intelligent Systems 5, no. 4 (2014): 48-56.

Kumar, Sushil, and Mousmi A. Chaurasia. "Assessment on stylometry for multilingual manuscript." IOSR Journal of Engineering 2, no. 9 (2012): 1-6.

Ouamour, Siham, and Halim Sayoud. "Authorship attribution of ancient texts written by ten arabic travelers using a smo-svm classifier." In 2012 International Conference on Communications and Information Technology (ICCIT), pp. 44-47. IEEE, 2012.

Ouamour, Siham, and Halim Sayoud. "Authorship attribution of short historical arabic texts based on lexical features." In 2013 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, pp. 144-147. IEEE, 2013.

Shaker, Kareem, and David Corne. "Authorship attribution in arabic using a hybrid of evolutionary search and linear discriminant analysis." In 2010 UK Workshop on Computational Intelligence (UKCI), pp. 1-6. IEEE, 2010.