Writing Styles of Salwa and Al-Qarni

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

This paper follows a recent court case which judged that "Don't despair" (لا تيأس) by Aaidh ibn Abdullah Al-Qarni was plagiarized from Salwa Aladian's "That is how they defeat Despair" (هكذا هزموا اليأس). We use techniques of computational stylometry, Hierarchical Agglomerative Clustering Analysis and Principal Components Analysis, to show that the disputed sections not only resemble Salwa's work in content, but also in writing style.

1 Introduction

Aaidh ibn Abdullah al-Qarni was born in 1960 in Saudi Arabia. He graduated from Imam Muhammad Ibn Saud University and became one of the most respected scholars in his country. Later he published more than 80 books in a very short time and became a very famous writer. In 2003 he introduced his book "Don't be sad" (لا تحزن) and it was one of the most successful books not only in Saudi Arabia but also in the world. The book addressed both Muslims and non-Muslims and was translated into many different languages. More than 10 million copies of his book were sold, and his followers increased. In 2012 Al-Qarni was found guilty of plagiarism, as Salwa Aladian claimed that he took many "Success stories" from her book "That is how they defeat Despair" (هكذا هزموا اليأس) and used them in his book "Don't despair" (لا تيأس). without putting any referencing to her work. Salwa stated that she collected these stories and used them in her book using her writing style. She said that Al-Qarni took the stories ready from her book and had stolen her book's introduction as well.

Al-Qarni's followers started to attack Salwa on social media and a fight of words took place between Salwa's family and Al-Qarni's supporters. Abdallah (2019) said that "The cleric had used his religious standing and media exposure to rally a group of dedicated students and followers against Salwa. These followers promptly used online forums and social media websites to attack the young author." The case took about one year, and Al-Qarni continued to deny the fact that he took the stories from Salwa's book. Salwa said that she trusted the justice in Saudi Arabia, and she introduced her case to the court. After reviewing the books, the court found that Al-Qarni was guilty and fined him. Al-Qarni paid 300,000 Saudi Riyals to Salwa and apologized to her after he stated that this was not his fault. Al-Qarni said that he had asked one of his students to collect stories about successful people and the student collected all of them from Salwa's book. In 2018, Al-Qarni was found guilty in another academic dishonesty case. The London-based Arabic language newspaper Arabi21 said that the heirs of the Syrian writer Abdel Rahman Raafat Pasha had won a case against Al-Qarni for allegedly stealing their father's book "Pictures of the Lives of the Companions." (See: https://m.arabi21.com/Story/1072807). Al-Qarni was convicted of reading specific paragraphs from the book on a television program without giving any reference to the original author. Al-Qarni had to pay a fine of 30,000 Saudi Riyals for infringing intellectual property rights and 120,000

Saudi Riyals to Abdel Rahman Raafat's family, in addition to the obligation to stop broadcasting or rebroadcasting the program. "After five years of litigation, we got a verdict against one of the preachers on behalf of a group of heirs. The preacher raided a book by their father, may Allah have mercy on him, and the judge sentenced him to a fine of 30 thousand, and ordered compensation of our client with 120 thousand, and we will ask for more than this" said lawyer Abdul Rahman Al-Lahim, who was appointed by the heirs of Pasha.

2 Related Work

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 bigram and rare words. The texts were collected from ten different authors who wrote their texts in the domain of travel. Sayoud (2012) also studied the difference in writing style between the Quran (The Muslim holy book) and the Hadith (sayings of the Prophet Muhammad). In his experiments, he extracted four segments from each book and used a hierarchical clustering algorithm to cluster the text according to style. For each book, four segments were extracted. In the first experiment, Sayoud investigated the use of discriminative words such as 'those' (الذين) and the word 'earth' (الأرض). He noticed that these words appeared more often in one of the books than in the other, so he decided to use them as a feature set. In the second experiment, he used word length to discriminate between the two styles. Finally, in the third experiment, he used a new parameter called COST. The COST parameter is a cumulative distance measuring the similarity between the ending of one sentence and the ending of the next. This gives an estimation measure of the poetic form of the text. When Arabic poets write a series of poems, they make a termination similarity between the neighboring sentences of the poem, such as the final syllable or letter. This known in Arabic as rhythm or "Qafeia".

Hadjadj and Sayoud (2016) also investigated the authorship of the Quran and the Hadith, implementing two experiments to explore whether their writing styles are similar of different. The first experiments used Manhattan centroid distance and the SMO-SVM classifier, while the second experiment used hierarchical agglomerative clustering. Three main features were extracted from the dataset: interrogative words, the discriminative words, and COST. The purpose of Sayoud's experiments on the Quran and Hadiths was to challenge the assumption that the Quran was invented by the Prophet Mohammed (and therefore not handed to him by God). He tried to show that the books have two distinctive styles and therefore could not have been written by the same author (Sayoud, 2017).

Alwajeeh, Al-Ayyoub, and Hmeidi (2014) manually collected texts from Arabic news websites. The texts consisted of 500 different articles written by five different authors. They then ran their data through two well-known classifiers, i.e. Naïve Bayes and SVM. In their experiment, they achieved near-perfect accuracy for both classifiers. AbdulRazzaq and Mustafa (2014) claim to be the first to the use classic Delta distance (Burrows, 2002), a measure of difference between two texts, to find the authorship of Arabic texts. In their study, they demonstrated the suitability of this method for Arabic texts by using a database containing 30 books written by five different authors. The results showed that word bigrams and word trigrams were the most suitable features for Arabic authorship studies. Rabab'ah *et al.,* (2016) used two common approaches, i.e. BOW and Stylometry Features (SF), to find authorship in Arabic tweets. The authors collected tweets from twelve famous Arabic Twitter users, professionals working in different fields, e.g. religion, politics, sport, academia, and music, each with many followers. Some of the features were extracted by the morphological analysis tool MADAMIRA (Pasha *et al.*, 2014). This tool returns useful information about the words like aspect, gender, mood, and part of speech. Other features like the unigram and BOW were extracted using the Weka tool (Hall *et al.*, 2009). The following classifiers were tested to find which set of features produced the highest accuracy: Naïve Bayes, Decision trees, and SVM. The results show that combining all the feature sets they computed yields the best result.

Shrestha et al. (2017) used Convolutional Neural Networks (CNNs) to perform authorship attribution task of tweets. They used character n-grams as the feature set and provided a strategy to improve model interpretability by estimating the importance of input text fragments in the predicted classification. The results showed that CNNs outperformed the previous methods.

3 Corpus Description

To build the experimental corpus we used some sample texts from Salwa's book "That is how they defeat the Despair" (هكذا هزموا اليأس) and sample texts from Al-Qarni's books "Thirty reasons for happiness" (ثلاثون سبب للسعادة) and "Characters from the Holy Quran" (شخصيات من القرآن الكريم). In addition, four samples from the disputed text were taken from the document produced by Salwa to compare the plagiarised text with her book. The length of each sample in the corpus was 2000 words. Table 1 shows the texts which were used in the experiments:

Text	Book	Author
X_1_1	The disputed text	?
X_2_1	The disputed text	?
X_3_1	The disputed text	?
X_4_1	The disputed text	?
Q_1_10_15_1[1]	Thirty reasons for happiness	Al-Qarni
Q_1_3_8_1	Thirty reasons for happiness	Al-Qarni
Q_2_30_40_1	Characters from the Holy Quran	Al-Qarni
Q_2_15_25_1	Characters from the Holy Quran	Al-Qarni
Q_2_3_9_1	Characters from the Holy Quran	Al-Qarni
S_30_40_1[2]	That is how they defeat the despair	Salwa
S_45_56_1	That is how they defeat the despair	Salwa
S_15_25_1	That is how they defeat the despair	Salwa
S_60_71_1	That is how they defeat the despair	Salwa
S_90_100_1	That is how they defeat the despair	Salwa

Table 1: Corpus Description.

In these experiments, we used Principal Components Analysis (PCA) to find whether the disputed texts would cluster with Al-Qarni's texts or with Salwa's text. We also used cluster analysis using the hierarchical agglomerative algorithm and the classic Delta intertextual distance measurement to cluster the texts according to textual similarity which is a proxy for writing style.

3.1 PCA Analysis Using the Most Frequent Words

Principal component analysis (Everitt, 2006) is a feature extraction technique. Sets of features (such as most frequent words) tend to co-occur in similar documents, and together they make up a principal component. This technique can be used to reduce the dimensionality of many variables by ranking and sequentially extracting the components according to how much they contribute to the overall variance in the model. The features which show up more in a specific group of texts and show up less in another group are used to discriminate between the texts. PCA is very useful when we have little data, and we had few texts to compare between Salwa's style and Al-Qarni's style. It would have been better if we could have used all the texts from the book La-Tayaas, but the decision of the court made it very difficult to find the whole book. Thus we used the texts which were produced by Salwa to compare her book with Al-Qarni's book and used the PCA technique to extract the most important features from these texts. The initial feature set which we used to discriminate between the two authors was the most frequent words. Figure 1 shows that the disputed texts which were represented by X_1, X2_1, X3_1, and X4_1 were placed on the right-hand side together with Salwa's samples. It is possible to plot the texts and the words which most characterize them on the same graph. Words which occur frequently in the texts are plotted near those texts, and those which

1 Q_1_10_15_1 means the sample was taken from the first book of Alqarni pages from 10 to 15

2 S_30_40_1 means the sample was taken from Salwa's book pages from 30 to 40

occur infrequently in those texts are plotted far away. The axes show how correlated the texts and words are with each principal component. In figure 1, the most useful extracted words are shown together with the different samples. Figure 1 shows that the writing styles of Salwa and Al-Qarni can be distinguished by the words of the first principal component, since Al-Qarni's texts appear on the left, and Salwa's texts (including the disputed samples) appear on the right. The following list of words was seen on Salwa's side:

مع: With / قبل: Before /أنه: Which is / لم: Not/ غير: Except /في: In / إليه: To / بين: Between /كل: All /التي: Which was/ كانت: Was / لها: For it

This list of words was seen on Alqarni's side: إن: So / إذا: If / أو: Or / عند: Have / فلا: Not/ لك: For you/ علي: On/ الي: To/ هذا: This/ لا: Not / بها: On that/ كيف: How/ هو: He/ يا: Oh /إلا: Except/ وما: And not/ وقال: Said/ يقول: Is saying/ لما: When.

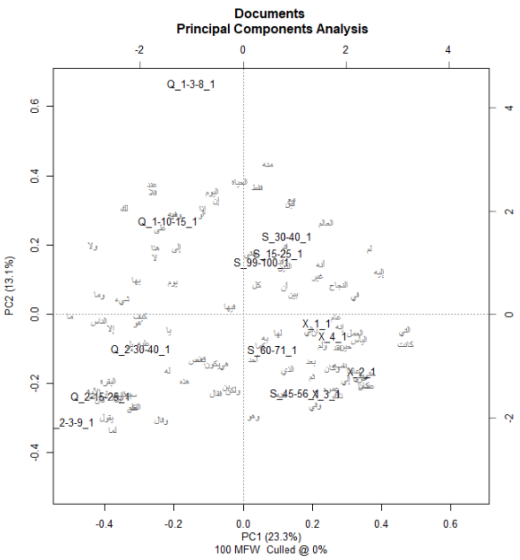

Figure 1: PCA including the MFW

3.2 PCA Using Morphemes

In this experiment we used the Farasa tool (Abdelali *et al.,* 2016) to extract the different morphemes contained in words as the feature set to discriminate between the two authors. For example the following morphemes (ب/ ت/ ها / ه) were observed on Salwa's side on the graph, while the morphemes (وا/ ل/ ف / ون) were on Alqarni's side. The following PCA graph (figure 3) shows the results obtained using this feature set. Once again, the disputed texts appeared on Salwa's side of the graph, showing that they were written in her writing style.

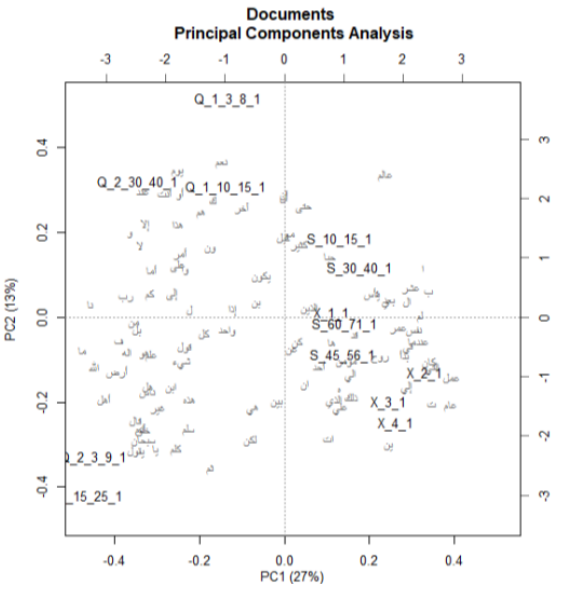

Figure 2 PCA using morphemes

3.3 Cluster Analysis Using the Most Frequent Words

In this experiment we used the Hierarchical Agglomerative Clustering Algorithm (HACA) (Everitt, 2005) to cluster the texts according to style. The most frequent words were used as a feature set and the classic Delta measure was used to measure the distance between the different texts. HACA displays the texts under analysis in a form of upside-down tree called a dendrogram, where the leaves are the texts and the branches show the distances between them. Thus, texts in a similar writing style will be placed close together, and dissimilar texts will be placed far apart. From figure 4, it is clear that the disputed tests X_1, X_2, X3_, and X_4 were clustered under the same branch which contained Salwa's samples. In addition to that, the samples were mixed with Salwa's samples, and they did not form a subset group. The samples which were taken from Al-Qarni's books were clustered together, and as we can see the samples from the first book formed a subset group and the samples from the second book also did so. This indicates that the stories of the "successful people" were taken from the book together with Salwa's style and very little paraphrasing was done for the texts. Salwa stated that she collected these stories and wrote them using her style to motivate the readers, and Al-Qarni took her effort without even putting any reference.

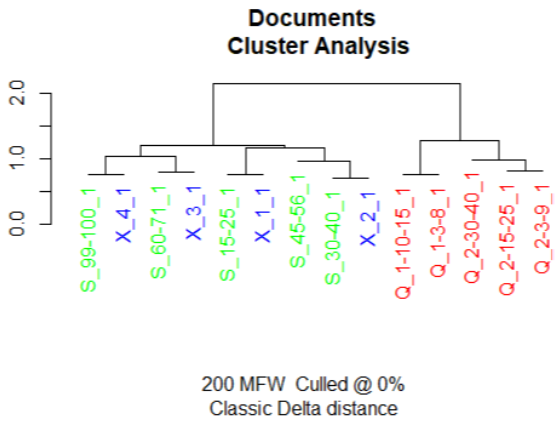

Figure 3: Cluster analysis using MFW

4 Conclusion

In this paper we introduced a recent case which occurred between two Saudi writers. Al-Qarni who is a very famous writer in Saudi Arabia was found guilty after the young female writer Salwa introduced a case against him to the court. To investigate Salwa's claim we collected some texts written by Al-Qarni and others written by Salwa to find which texts the disputed texts were more similar to. We used different features including single words, and morphemes contained in words. In addition, two different methods were used to do the analysis namely Hierarchical Agglomerative Clustering Analysis and Principal Component Analysis.

To sum up, Al-Qarni confirmed that the stories were taken from Salwa's book as the designated student for the task of collecting them took all the stories from one source which was Salwa's book. This was a problem mentioned by Al-Qarni himself, but another problem was that, as we can see from the results above, the collected stories were included in Al-Qarni's book without doing more paraphrasing to reproduce the stories in Al-Qarni's writing style. This made Salwa's fingerprint still visible in the texts, as we saw when the disputed texts clustered together with the texts in Salwa's branch of the HACA dentrogram, and the disputed texts appeared on Salwa's side of the PCA plot.

References

Abbasi, Ahmed, and Hsinchun Chen. "Applying authorship analysis to extremist-group web forum messages." IEEE Intelligent Systems 20, no. 5 (2005): 67-75.

Abdallah, Mariam. 2019. Academic Theft!. [online] Muftisays.com. Available at: https://www.muftisays.com/forums/77-taqleed--the-straight-path/9852-academic-theft.html [Accessed 27 Feb. 2019].

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.

AbdulRazzaq, Ammar Adil, and Tareef Kamil Mustafa. "Burrows-Delta Method Fitness for Arabic Text Authorship Stylometric Detection." (2014).

Albadarneh, Jafar, Bashar Talafha, Mahmoud Al-Ayyoub, Belal Zaqaibeh, Mohammad Al-Smadi, Yaser Jararweh, and Elhadj Benkhelifa. "Using big data analytics for authorship authentication of arabic tweets." In Proceedings of the 8th International Conference on Utility and Cloud Computing, pp. 448-452. IEEE Press, 2015.

Al-Qarni, Aaidh. "Don't be sad." Saudi Arabia: International Islamic Publishing House (2003).

Burrows, John. "'Delta': a measure of stylistic difference and a guide to likely authorship." Literary and linguistic computing 17, no. 3 (2002): 267-287.

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

El-Fiqi, Heba, Eleni Petraki, and Hussein A. Abbass. "A computational linguistic approach for the identification of translator stylometry using Arabic-English text." In 2011 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2011), pp. 2039-2045. IEEE, 2011.

Everitt, Brian S. An R and S-PLUS® companion to multivariate analysis. Springer Science & Business Media, 2006.

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.

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.

Pasha, Arfath, Mohamed Al-Badrashiny, Mona T. Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan

Roth. "Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of arabic." In LREC, vol. 14, pp. 1094-1101. 2014.

Rabab'ah, Abdullateef, Mahmoud Al-Ayyoub, Yaser Jararweh, and Monther Aldwairi. "Authorship attribution of Arabic tweets." In 2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA), pp. 1-6. IEEE, 2016.

Sayoud, Halim. "Authorship classification of two old arabic religious books based on a hierarchical clustering." In Workshop Organizers, p. 65. 2012.

Sayoud, Halim. "AUTHORSHIP DISCRIMINATION ON QURAN AND HADITH USING DISCRIMINATIVE LEAVE-ONE-OUT CLASSIFICATION." (2017).

Shrestha, Prasha, Sebastian Sierra, Fabio Gonzalez, Manuel Montes, Paolo Rosso, and Thamar Solorio. "Convolutional neural networks for authorship attribution of short texts." In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, pp. 669-674. 2017.