Distance-Based Authorship Verification Across Modern Standard Arabic Genres

Hossam Ahmed
Leiden University Institute for Area Studies
Witte Singel 25, 2311 BZ, Leiden, The Netherlands
h.i.a.a.ahmed@hum.leidenuniv.nl

Abstract

Authorship Verification (AV) is a type of stylometric analysis that addresses an authorship problem where, given a document of unknown origin and a set of documents written by a known author, the task is to identify whether the document is indeed written by that author. Previous research uses a number of techniques to address this problem. Most successful techniques in Classical Arabic as well as other languages use an SVM method supported by a distance measure in vector space and a distance/similarity threshold for accepting the document as authentic. While Arabic Authorship Attribution (where the task is to attribute the question document to one of several candidates) surveys and evaluates the usability of different distance measures, this paper is the first to provide such overview for Modern Arabic AV. Using a corpus of short texts from five common Modern Standard Arabic genres, this paper evaluates four common distance measures (Canberra, Manhattan, Cosine, and Jaccard) with a number of lexical, syntactic, and morphological features. The results show that Canberra Distance is a best performing distance measure in most genres, with an accuracy rate of up to 97.8%, well over highest known baseline.

1 Introduction

This paper compares the accuracy of Authorship Verification (AV) in five Modern Standard Arabic (MSA) genres using four popular distance measures: Manhattan Distance, Canberra Distance, Cosine Distance, and Jaccard distance. The genres in question are fiction and non-fiction books, and articles on economics, politics, and newspaper columns.

Authorship Verification (AV) is a type of authorship analysis problem that addresses the question of whether a question document is written by a known author, given a corpus of authentic documents known to be written by that author. AV is often compared to Authorship Attribution (AA), where there is a set of known candidate authors, and the task is to determine which one of them is the author of the question document. Both AV and AA are relevant in the areas of corpus linguistics, stylistic and literary analysis, Digital Humanities, and forensic linguistics.

This paper is organized as follows: section 2 gives an overview of relevant literature and outlines the research question. Section 3 describes the corpus used and feature extraction. Section 4 outlines the Authorship Verification method and distance measured used in the experiments. The results are described and discussed in sections 5 and 6.

2 Related Work

When approaching AA and AV as Machine Learning (ML) tasks, AV differs essentially from AA in that the former involves only positive training data (a corpus known to be written by just one author). AA, on the other hand, involves a set of documents for each of the candidate authors. It can be argued that an AA task is easier, in the sense that all is needed is to determine which corpus is most similar to the question document. In AV, the alternative corpus is virtually that of any other author.

2.1 Arabic Authorship Attribution

AA is often approached as a classification problem. Literature on AA is extensive. For Arabic ML-based research, there has been much progress. Abbasi & Chen (2005a, 2005b) use an elaborate combination of C4.5 and SVM classifiers, combined with an ensemble of linguistic and non-

linguistic features to analyze web and forum authorship. They find that SVM outperforms decision trees in AA, reaching accuracy of 94% for Arabic. SVM has also been used with a number of features in other AA contexts with success. Ouamour & Sayoud (2013) achieve 80% accuracy using Rare Words as the SVM feature of choice. Howedi & Mohd (2014) show that a small training set can render high AA accuracy using Naïve Bayes Bag of Words (96.67%) and SVM using character or word tetragram (93.33%). For modern Arabic, Altakrori et al. (2018) investigate AA in Twitter posts using an array of n-gram character, word, and sentence features, to be used with a number of ML algorithms (Naïve Bayes, SVM, Decision Trees, and Random Forests). For tweets, Random Forests seems to outperform other approaches.

2.2 Arabic Authorship Verification

AV tasks are often seen as AA tasks with the added complication that there is only one author to consider. A reasonable and successful approach in AA is to build a profile of certain features for each of the given authors and the question document, compare the profile of the question document to each of the author profiles, and make a decision using any of the approaches outlined in the previous section. This approach is not immediately accessible to AV because there is only one available dataset to create a profile; that of a single author. Two main approaches emerged to overcome this obstacle. The Imposters Method supplements the training data with a corpus of distractors, text known to be written by other authors, converting the task to an AA problem, and using familiar AA techniques. An example of such approach in Arabic can be seen in Arabic Twitter posts (Altakrori et al., 2018) where the stated context of the approach is law enforcement, where the authenticity of a tweet is needed as evidence. The authors frame the problem as attributing a question tweet to one of a number of suspects. This method suffers from two main drawbacks. First, using the Imposters method incurs additional computational cost as multiple profiles will be created. Second, the quality of AV prediction relies, at least partially, on the selection of the supplementary corpus. The perpetrator in the tweets example (ibid) may not be one of the usual suspects – the attribution problem will then return the suspect with the closes style out of the group provided. Similar issues arise in literary analysis contexts.

The Author Profiling method aims at avoiding the problems that arise with the Imposters method. Within this method, features are extracted from the known corpus and suspect document to create author profiles. If the two profiles match or are sufficiently similar, the document is deemed authentic, otherwise it is judged to be written by another author. Determining similarity and deciding on the threshold for acceptance are key questions in this approach. In languages other than Arabic, Halvani, Winter, & Pflug (2016) use an ensemble of n-gram features over 5 European languages using SVM-calculated distance metric based on Manhattan Distance (Burrows, 2002). They determine the similarity threshold of acceptance (θ) through Equal Error Rate (EER), a point where false negatives and false positives in the training data are equal. They achieve accuracy rates in the mid-70% range, depending on the language tested. It can be seen in this example that negative training data is still needed. EER is also used with a distance based metric based on compression models rather than linguistic features (Halvani et al., 2017), also relying on negative training data to determine θ, with remarkable improvement in processing time, yet slightly lower accuracy than best-performing approaches. Jankowska, Milios, & Kešelj (2014) define θ in terms of the maximum dissimilarity within the training set, completely dispensing with negative data in training. Using common character n-grams and Nearest Neighbor technique, this technique achieves accuracies in the high 80% when applied to the English, Spanish, and Greek datasets from PAN-2013 (Stamatatos et al., 2014). Benzebouchi et. al (2018) use word embeddings and a voting system between SVM and NN techniques to produce high-accuracy AV.

There is little research on Arabic AV. In Classical Arabic, Ahmed (2018) uses an author profiling technique, a similarity metric based on Burrows (2002), and defines θ in terms of simple Gaussian technique to show that stem bigrams offer best accuracy performance (87%) for Classical Arabic. There is no research that deals with MSA. Furthermore, it is not immediately clear if the similarity metric (based on Manhattan Distance) is also optimal in non-literary genres. A

comparison of the effectiveness of different distance measures is not available for Arabic AV.

2.3 Research Question

The research outlined above indicate that the careful choice of classifier and relevant feature sets contributes to better AA and AV accuracy. Genre distance metric also seem to play a role in AA. García-Barrero, Feria, & Turell (2013) show that AA accuracy is sensitive to genre, even in closely related genres (literary criticism and short stories). Ouamour & Sayoud (2018) conduct a broad survey of distances and feature sets used in Arabic AA, showing that Manhattan centroid gives highest average accuracy in Arabic AA. The effect of distance or genre has not been studied in Arabic AV.

This is the first study to look at the effect of distance measures and feature selection in modern Arabic Authorship Verification. This paper addresses the following questions:

1. Does feature selection affect the accuracy of AV across MSA genres?
2. Does distance measure selection affect the accuracy of AV across MSA genres?

Depending on the feature set under investigation, the first question addresses lexical, grammatical, and stylistic characteristics of an individual writer, but also of the genre under discussion. The second question addresses the role of feature frequency in the success of AV in Arabic.

To answer these questions, this paper reports the results of a number of experiments examining the accuracy of distance-based AV in modern Arabic in five genres: opinion columns, economics, politics, fiction and nonfiction. The paper compares the accuracy of best performing features in the survey conducted for AA by Ouamour & Sayoud (2018): Manhattan Distance, Canberra Distance, Jaccard Distance, and cosine similarity. The feature set and similarity threshold θ used in this paper are similar to those used by Ahmed (2017, 2018), as they report highest accuracies for Classical Arabic AV. Specifically, this paper will use n-grams of tokens, stems, trilateral roots, and part-of-speech tags.

3 Corpus used

A total of 125 documents from five common genres in Modern Standard Arabic are selected as follows. Five authors are selected from each genre. For each author, five documents are collected.

Author	Source
Fiction	Hindawi Foundation book repository www.hindawi.org
Ali Al-Jaarim	
Abdul Aziz Baraka Sakin	
Nicola Haddaad	
Nawaal Al-Saadaawi	
Georgi Zidaan	
Non-fiction	
Abbas Al-Aqqaad	
Ismail Mazhar	
Salama Moussa	
Fouad Zakareyya	
Zaki Naguib Mahmoud	
Economics	www.almasryalyoum.com
Musbah Qutb	
Mohammed Abd Elaal	
Bissan Kassab	www.madamasr.com
Waad Ahmed	
Yumn Hamaqi	www.ik.ahram.org.eg
Politics	
Alaa Al-Aswani	www.dw.com
Wael Al-Semari	www.youm7.com
Danadarawy Al-Hawari	
Belal Fadl	www.alaraby.co.uk
Salma Hussein	www.shorouknews.com
Opinion columns	www.shorouknews.com
Ashraf Al-Barbari	
Emad Eldin Hussein	
Fatima Ramadan	
Mostafa Kamel El Sayyed	
Sara Khorshid	

Table 1: Corpus used.

Table 1 lists the authors and source of the documents used for the corpus. Whenever possible, authors and texts are selected from similar backgrounds e.g. Egyptian writers or Egyptian web sites, to minimize the effect of language variation across dialects. The corpus is collected from same source for each genre whenever possible to minimize any potential editorial effect.

Domain	Avg. size
Opinion columns	746
Economics	765
Fiction	1010
Nonfiction	1001
Politics	760

Table 2: Average document length (tokens).

3.1 Preprocessing and feature extraction

For Economics, politics, and opinion columns, documents are downloaded as text-only (UTF-8) documents. Titles, by-lines, and other front matter are removed. For fiction and non-fiction, documents collected are entire books in e-book (epub) format. They are converted to plain text (UTF-8), then sampled by using about 1100 words from the middle of the book using regular expressions delimited by space, to avoid material that may be repeated verbatim for a given author (front matter, acknowledgement, repeated preface, dedication, etc.). Punctuation and non-Arabic characters are removed. Table 2 shows average document length per genre after pre-processing.

The feature token is taken to represent Arabic words, and is defined as a sequence of Arabic characters separated by white space (note that non-Arabic characters, digits, and punctuation marks have been removed in preprocessing). The pre-processed text is passed through MADAMIRA version 2.1 (Pasha et al., 2014) with standard settings. Part-of-speech (POS) tags and word stems are then extracted from the analysis produced by MADAMIRA. Roots are extracted from the plain-text corpus using ISRI Stemmer in NLTK (Bird et al., 2009). Table 3 shows an example of features extracted from the pre-processed word 'المؤلفين'.

Preprocessed word	المؤلفين
Token	المؤلفين
Root	ألف
Stem	مؤلف
POS tag	noun

Table 3: Example of features extracted from an input word.

4 Verification method

This section outlines the verification method of the experiment.

AV problem: the authorship problem is defined as $p(D_u, D_A) \rightarrow \{1, 0\}$ where D_u is a document of questionable attribution to an author A, and $D_A = \{D_{A,1}, D_{A,2}, \ldots\}$ is the set of documents of known attribution to A. As this is an AV, rather than AA, problem, D_A is of a single author, and there is only one set per problem. The AV procedure should return 1 if D_u is written by A and 0 if not. No 'unknown' response is allowed.

Data representation: simplifying the problem, all the known documents in D_A are concatenated to create a single document.

Feature engineering: D_A is a document with sequence of tokens, roots, POS tags, or stems produced by preprocessing. N-grams of relevant features are created, where $n \in \{1, 2, 3, 4\}$. The known and question documents are vectorized over term frequencies of the relevant feature n-grams using Scikit-Learn (Pedregosa et al., 2011).

Computing distance metrics: Four distance metrics are calculated between D_u and D_A. based on Ouamour & Sayoud (2018) the four distance measures are Manhattan Distance, Canberra Distance, Cosine Distance, and Jaccard distance. Stamatatos Distance is not implemented, as it performs consistently poorly in their survey.

Manhattan Distance: for unknown document D_u, known corpus D_A, and normalized frequency of feature f n-gram, Manhattan Distance is calculated as:

$$Man(D_u, D_A) = \sqrt{\sum_{f=1}^{n} |D_{u,f} - D_{A,f}|}$$

Canberra Distance is calculated as

$$Can(D_u, D_A) = \sum_{f=1}^{n} \frac{|D_{u,f} - D_{A,f}|}{|D_{u,f}| - |D_{A,f}|}$$

Cosine Distance is defined as

$$CosDist(D_u, D_A) = \frac{D_{u,f} . D_{A,f}}{||D_{u,f}||_2 - ||D_{A,f}||_2}$$

Jaccard Distance is defined as

$$Jacc(D_u, D_A) = \frac{|D_{u,f} \cap D_{A,f}|}{|D_{u,f} \cup D_{A,f}|}$$

Threshold determination: The training phase of this method is comprised of calculating a similarity threshold θ above which D_u is considered authentic. following Ahmed (, 2017), the acceptance threshold θ is dynamically calculated for each $D_A = \{D_{A,1}, D_{A,2}, \ldots\}$ by calculating the distance for each known document k and the rest of the known documents:

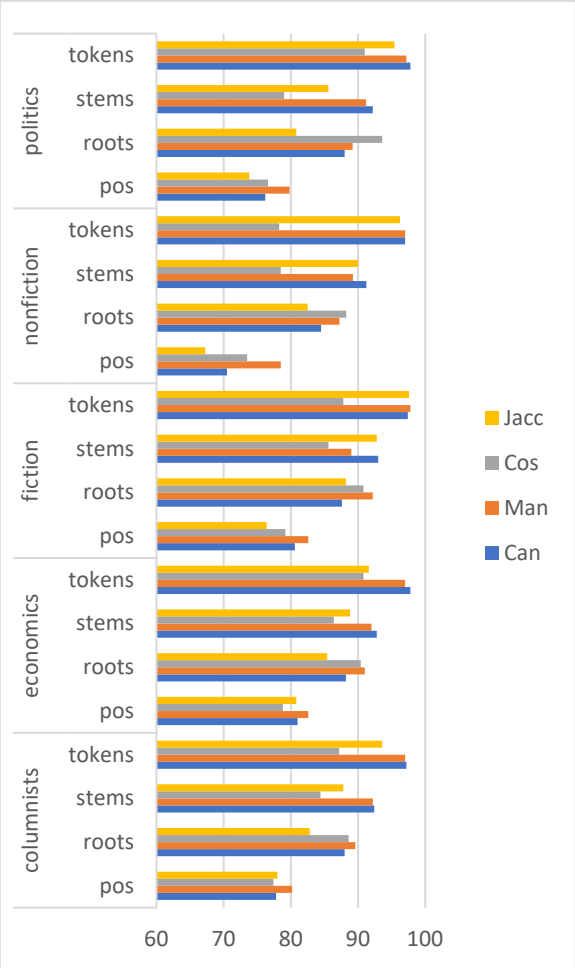

Figure 1: Distance accuracies per genre (unigrams).

$$Dist_k = Dist(D_k, D_{A-k})$$

θ is then defined as the lower bound of the confidence interval of the values of all members of D_A at p = 0.005.

Verification: The testing phase consists of calculating the distance for each document in a given genre against the known corpus for each author. Training and testing data come from the same genre. The document is considered unauthentic if distance $Dist(D_u, D_A) > \theta$ and authentic otherwise.

Evaluation and Baseline: Evaluation of the results is done through the leave-one-out method. Accuracy is defined as follows:

$$accuracy = \frac{Correct\ predictions}{Total\ predictions}$$

The baseline accuracy for this experiment is that used by Ahmed (, 2018) using Manhattan Distance in Classical Arabic and the same θ used in this paper. The best performing feature ensemble for the baseline is stem bigrams.

Domain	**Distance measure**	**Accuracy**
Opinion columns	Canberra	97.2%
Economics	Canberra	97.8%
Fiction	Manhattan	97.8%
Nonfiction	Manhattan, Canberra	97%
Politics	Canberra	97.8%
Baseline		**87.1%**

Table 4: Best performing feature/distance measure per domain.

5 Results

The testing method returned results for all genres that are consistently and considerably above the baseline reported for Classical Arabic. For all genres, the best performing feature is token unigrams, with accuracy ≥ 97%, albeit with some variation in the winning distance measure. Table 4 shows the best performing distance measure per genre.

Figure 1 shows distance accuracies per feature unigram over genres. The figure shows that in four out of the five genres, Canberra Distance is the best performing distance measure to be used with the tested method, with Manhattan Distance coming at a close second. Cosine distance and Jaccard distance perform considerably less accurately, although their best performance is still consistently higher than the baseline.

Another finding of the experiments is that higher n-gram feature assemblies perform worse than their unigram counterparts to varying degrees. Figure 2 compares distance measure accuracies across various n-grams. It shows that for unigrams, the distance measures perform at higher 90% accuracies, while for n = 2 – 4, accuracies drop to mid- and low-80%.

6 Discussion

The overall trend of the results – as far as the research question of this paper is concerned – is expected. AV accuracy is sensitive to frequencies across genres. Overall, distance measures that are least sensitive to frequency (Jaccard distance and cosine distance) underperform compared to those which incorporate frequency (Canberra,

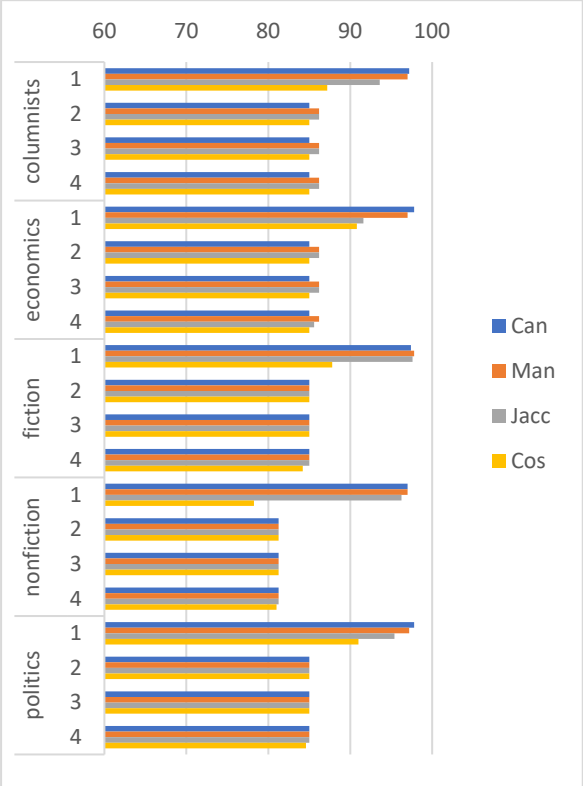

Figure 2: Distance n-gram accuracies per genre.

Manhattan). The slightly improved accuracies delivered by Canberra Distance over Manhattan Distance across all five genres reflects the value of weighing the less common terms (in this case tokens), as Canberra Distance is more sensitive to vectorized values of smaller values than Manhattan Distance.

Some of the unexpected results pertain to the best performing feature, and the improvement of accuracies in this experiment over best known baseline in Classical Arabic. Best performance in this experiment is at least 10% higher accuracy than reported by Ahmed (, 2018). Using the same feature ensemble and distance measures reported for best results in that reference (stem bigrams and Manhattan-distance based similarity) renders accuracies slightly lower than the Classical Arabic data (80% - 85% MSA, depending on genre, compared to 87.1%, and token unigrams are 20% less accurate than results reported here). This difference might be attributed to stylistic variation, change in language convention (higher reliance on loan-words or some similar lexical factor that MSA uses to allow writers to distinguish themselves, while in CA innovation might be said to be at a deeper lexical level). Still, the difference in best-performance is very high. It can be explained in terms of size effects. The Author (Ahmed, 2018) notes that the size of the documents used is very large, and that there is no gain in performance after using more than 1% of the corpus used, and alludes that using even smaller corpora might help improve predictions. While CA texts are volumes in size, the texts used in this experiment are less than 1100 tokens long. Another possibility is the difference in calculating the distance. Using Manhattan Distance and Canberra Distance in their raw form in this experiment causes a tighter cluster, smaller distances, than used to generate the baseline (through the square root or division over separate frequencies). The baseline uses 'delta;' a distance measure based on Manhattan Distance, but does not take the square root (Ahmed, 2017; Burrows, 2002). This means that during the training phase, known documents will generate similarity values that are more spread over the vector space, and a less tight confidence interval for calculating θ.

A related point of difference to existing literature is that the best performing feature in this experiment is Canberra Distance, which ranked low in Classical Arabic AA survey (Ouamour & Sayoud, 2018). This difference can be an additional indicator that MSA differs stylistically from Classical Arabic, note that the discussion above for Arabic AV also compares this work to Classical Arabic. It could also be related to the different nature of the task (AA vs. AV).

Another unexpected finding is consistency across genres. One would expect that authors in different genres would differentiate themselves differently. For example, a genre like novels (fiction) or opinion columns would be expected to give authors more latitude to differentiate themselves by using more varied phrase structures than, say, economics. This in theory would reflect in better differentiation through features such as POS n-grams. However, this does not seem to be the case, and lexical selection is consistently the differentiating factor across the five genres under discussion. On the other hand, this is good news on the computational side; a simple Bag-of-Words, minimal preprocessing, and a simple similarity metric will yield excellent results in efficient computation time.

The superior performance of token unigrams raises a number of questions. The first issue is related to genre characteristics. In genres such as economics, politics, and opinion columns, it is

likely that texts go to post-editing prior to publication, ant this could affect certain features more than others. If an editor is more likely to change sentence phrasing and grammatical 'errors' than alter word choice, purely lexical features (tokens) would be a better reflection of the author's style than the stylesheet of the publisher. This, however, does not seem to be the case in the current experiment. Token unigrams are also the most effective feature ensemble in fiction and non-fiction, where post-editing is not expected. In opinion columns, the whole corpus is extracted from a single source, potentially reducing or neutralizing any possible effects of post-editing. Token unigrams are still the most effective feature ensemble.

The second question related to the higher performance of token unigrams comes from the nature of feature extraction. POS tags and stem features are extracted using MADAMIRA with standard settings and roots are generated using ISRI. MADAMIRA is reported to have 95.9% accuracy in POS tagging and 96.0% for stemming (Pasha et al., 2014) while ISRI reports recall and precision values of less than 48% (Taghva et al., 2005). Whether the development of better morphological analyzers could indeed reveal that the value of token unigrams in AV is overstated is an empirical question that I leave for future research.

7 Conclusion

In this paper I have shown that distance measures that are sensitive to term frequency deliver higher accuracies in AV tasks in MSA across five common genres. I have also shown that a simple BoW technique together with a simple non-negative-evidence algorithm that uses Canberra Distance to determine AV can deliver very high accuracies with minimum pre-processing.

Future research should focus on cross-domain AV. Would the same method and distance measures perform with the same behavior if the training set comes from a domain and the test document from another? The fact that tokens are the key features might affect that outcome. On the other hand, as Canberra Distance is weighted to be more sensitive to less common vectors, it may be likely that domain-specific tokens be not so influential as to affect the AV task. I leave this question to future research.

References

Ahmed Abbasi and Hsinchun Chen. 2005a. Applying Authorship Analysis to Arabic Web Content. In Paul Kantor, Gheorghe Muresan, Fred Roberts, DanielD. Zeng, Fei-Yue Wang, Hsinchun Chen, and RalphC. Merkle, editors, *Intelligence and Security Informatics SE - 15*, volume 3495 of *Lecture Notes in Computer Science*, pages 183–197. Springer Berlin Heidelberg.

Ahmed Abbasi and Hsinchun Chen. 2005b. Applying authorship analysis to extremist-group Web forum messages. *IEEE Intelligent Systems*, 20(5):67–75.

Hossam Ahmed. 2017. Dynamic Similarity Threshold in Authorship Verification: Evidence from Classical Arabic. *Procedia Computer Science*, 117:145–152.

Hossam Ahmed. 2018. The Role of Linguistic Feature Categories in Authorship Verification. *Procedia Computer Science*, 142:214–221.

Malik H. Altakrori, Benjamin C. M. Fung, Steven H. H. Ding, Abdallah Tubaishat, and Farkhund Iqbal. 2018. Arabic Authorship Attribution: An Extensive Study on Twitter Posts. *ACM Trans. Asian Low-Resour. Lang. Inf. Process*, 18(1):51.

Nacer Eddine Benzebouchi, Nabiha Azizi, Monther Aldwairi, and Nadir Farah. 2018. Multi-classifier system for authorship verification task using word embeddings. *2nd International Conference on Natural Language and Speech Processing, ICNLSP 2018*:1–6.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

John Burrows. 2002. "Delta": a measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3):267–287.

David García-Barrero, Manuel Feria, and Maria Teresa Turell. 2013. Using function words and punctuation marks in Arabic forensic authorship attribution. In Rui Sousa-Silva, Rita Faria, Núria Gavaldà, and Belinda Maia, editors, *Proceedings of the 3rd European Conference of the International Association of Forensic Linguists*, pages 42–56, Porto, Portugal. Faculdade de Letras da Universidade do Porto.

Oren Halvani, Christian Winter, and Lukas Graner. 2017. Authorship Verification based on Compression-Models.

Oren Halvani, Christian Winter, and Anika Pflug. 2016. Authorship verification for different languages, genres and topics. *Digital Investigation*, 16:S33–S43.

Fatma Howedi and Masnizah Mohd. 2014. Text Classification for Authorship Attribution Using Naive Bayes Classifier with Limited Training Data. *Computer Engineering and Intelligent Systems*, 5(4):48–56.

Magdalena Jankowska, Evangelos Milios, and Vlado Kešelj. 2014. Author Verification Using Common N-Gram Profiles of Text Documents. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics*:387–397.

Siham Ouamour and Halim Sayoud. 2013. Authorship Attribution of Short Historical Arabic Texts Based on Lexical Features. In *2013 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery*, pages 144–147.

Siham Ouamour and Halim Sayoud. 2018. A Comparative Survey of Authorship Attribution on Short Arabic Texts.

Arfath Pasha, Mohamed Al-badrashiny, Mona Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan M Roth. 2014. MADAMIRA : A Fast , Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. *Proceedings of the 9th Language Resources and Evaluation Conference (LREC'14)*:1094–1101.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, and Vincent Dubourg. 2011. Scikit-learn: Machine learning in Python. *Journal of machine learning research*, 12(Oct):2825–2830.

Efstathios Stamatatos, Walter Daelemans, Ben Verhoeven, Martin Potthast, Benno Stein, Patrick Juola, Miguel A. Sanchez-Perez, and Alberto Barrón-Cedeño. 2014. Overview of the author identification task at PAN 2014. In *CEUR Workshop Proceedings*, volume 1180, pages 877–897.

Kazem Taghva, Rania Elkhoury, and Jeffrey Coombs. 2005. Arabic stemming without a root dictionary. In *Information Technology: Coding and Computing, 2005. ITCC 2005. International Conference on*, volume 1, pages 152–157. IEEE.