Authorship Verification for Arabic Social Media Texts Using Arabic Knowledge-Base Model (AraKB)

Fatimah Alqahtani[1] and Helen Yannakoudakis[2]
Department of Informatics, King's College London, UK
Fatimah.alqahtani@kcl.ac.uk[1] , Helen.yannakoudakis@kcl.ac.uk[2]

Abstract

The issue of verifying authorship has been a controversial and much disputed subject within the field of digital forensics and cyber investigations. Although extensive research has been carried out on authorship verification tasks, few studies have analyzed Arabic social media texts. This paper seeks to overcome this limitation and presents a new knowledge-based model to enhance Natural Language Understanding and thereby improve authorship verification performance. The proposed model provided promising results that would benefit research for different Natural Language Processing tasks for Arabic.

1 Introduction

Arabic has a very rich vocabulary; each word has many derivatives that describe the root meaning in more specific or nuanced ways. For example, the word (thirst /عطش) has up to 45 synonyms, including those alluding to different stages of thirst. Compared to other languages, Arabic's structure increases the complexity of pre-processing, whereby unnecessary characters must be removed or carefully replaced. Moreover, the complexity of Arabic morphology tends to increase the set of features, syntax, and semantic structures, which might not be effective for the purposes of authorship verification (AV), the process of determining whether or not two pieces of writing are written by the same author by comparing their writing styles (Abbasi & Chen, 2005).

Although extensive research has been carried out on AV in different languages, few studies have focused on the Arabic language. AV in Arabic entails numerous particular linguistic difficulties, including with regard to inflection, elongation, diacritics, word length (Abbasi & Chen, 2005), and other challenges, as described below:

- Inflection: In Arabic, one word can generate three or more different words with minor change, therefore orthographical properties result in lexical variation. When the number of features increases, then determining the right number of features may impact authorship analysis performance (Larkey & Connel, 2001).
- Elongation: Proper pre-processing is necessary to remove unnecessary characters, but this may lose word emphasis or stylometric features of writers (Shaalan & Raza, 2009).
- Diacritics: Less effective when using word-based syntactic features (Abbasi & Chen, 2005).
- Word length: Shorter word lengths in Arabic (e.g., compared to English) reduces the effectiveness of lexical features (Abbasi & Chen, 2005).
- Diglossia: May not provide significant features or an adequate ontology to provide proper mappings (Badawi, 1996).
- Grammatical structure: Arabic dialect analysis creates more challenges.
- Capitalisation and punctuation: Identifying patterns is challenging (Ryding, 2005).
- Agglutinative constructs: Difficult parts of speech tagging could degrade the stylometric features of AV (Shaalan & Raza, 2009).

Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP), pages 205 - 213

In addition to the previous challenges, as AV mostly depends on the style of writing, minimal data pre-processing is required. Unlike most NLP tasks, such as sentiment analysis and text classification, AV problems cannot undergo extensive data pre-processing, as stemming, normalization, diacritics removal, and other pre-processing techniques would eliminate the author's style of writing, and therefore make AV more challenging. The challenge becomes even greater with authorship analysis tasks for very short texts (Azarbonyad, 2015; Luyckx & Daelemans, 2011), such as those on social media platforms. Hence, a minimal number of Arabic AV studies have been conducted due to the inherent difficulty of such undertakings.

Consequently, there is a need to investigate different linguistic features that could help to improve performance of Arabic AV for Arabic short texts (particularly Twitter posts). This paper presents a novel method and a new presentation of data to verify the authorship of Arabic texts specifically on Twitter; however, the experiment in principle could be applicable to other social media platforms and any short Arabic texts.

The following section presents a brief overview of the recent work on Arabic AV, then Section 3 explains the research methodology used in this work. Section 4 presents the experimental setup and results, and Section 5 discusses the results of the experiments. Finally, Section 6 summarizes the main conclusions and identifies areas for future investigation.

2 Related Work

As mentioned earlier, there are few studies of Arabic AV, and those which have been undertaken mainly analyzed very long texts, such as novels (Kumar & Chaurasia, 2012) and other books (Ahmed, 2017, 2018). In the following we will review the ones conducted on medium to short texts, as they are more relevant to the current study (which pertains to tweets).

An extensive set of documents was collected from Dar Al-Ifta,[1] consisting of 3,000 balanced datasets and 4,686 documents from unbalanced datasets (Al-Sarem, Emara, Cherif, Kissi, & Wahab, 2018). The method is based on the frequency-based features of unigrams, bigrams, and trigrams, and on style-based features (character, lexical, syntactic, semantic, content-specific, structural, and language-specific). First, the data were filtered, and TFIDF vectors were created. A bootstrap aggregating learner was then used to estimate the classification based on a maximum number of votes technique. Several stylometric and frequency-based features were used, showing that combining the bigram model with style-based features achieved the highest accuracy. However, it was unclear whether authors' documents were used in training or chunking in such lengthy article datasets.

Two experiments by (Ahmed, 2019a) sought to find the best feature ensemble, using the features of tokens, stems, root, diacritics, and POS tags of n-grams (1 to 4) as features for Arabic author verification. The author used a dataset consisting of 253 documents written by different authors from five domains. The average document sizes for the studied domains were 802 for columnists, 820 for economics, 1159 for fiction, 1108 for nonfiction, and 850 for politics. The accuracy for each domain varied from 80-84.53%. It is important to note that domains with the smallest sample size achieved the worst results. The second experiment was to find the effect of training or testing sample size, and it revealed that the training dataset size did not correlate with improved accuracy for the AV method. In conclusion, the study found that a training set with a smaller number of documents outperformed one with a larger number of documents.

Arabic AV using 125 documents from five common genres in Modern Standard Arabic (MSA), including opinion columns, economics, fiction, nonfiction, and politics, was undertaken by evaluating SVM-calculated distance metrics of the Canberra, Manhattan, Cosine, and Jaccard measures using tokens, stems, and POS tags as features (Ahmed, 2019b). It was found that the Canberra distance measure was the best-performing distance measure in most genres, with an accuracy rate as high as 97.8%. However, the method omits digits, punctuation marks, and special characters in pre-processing, which limits the applicability of these findings to short texts.

In our recent work on Arabic AV (Alqahtani & Dohler, 2022a), we collected a dataset consisting

[1] https://www.dar-alifta.org/Foreign/default.aspx

of 100 Twitter users written in the Arabic language, whereby each user had 1000 to 3000 tweets. Firstly, we extracted a number of sylometric (content-free) features compatible with both the Arabic language and with Twitter posts. Comparing different classifiers, we found that Gradient Boosting, with an average accuracy of 0.75, outperformed Random Forest, Support Vector Machine, and k-Nearest Neighbor. In the second experiment, the effect of combining content-specific features (e.g., TF-IDF) with the extracted stylometric features was tested, which improved the accuracy by almost 2%. The performance of using a combination of stylometric and TF-IDF resulted in 0.77 average accuracy and F1-score.

In general, it can be concluded that authorship analysis tasks depend on the feature set, the number of authors, and the dataset genres that reflect the Arabic language type (Classical, MSA, or Colloquial). In addition, the changing behavior of authors is an inherent problem that affects solving authorship verification problems.

The root cause of the limited number of works on Arabic-language authorship analysis is the inherent characteristics of the Arabic language itself. Compared to other languages, Arabic-language structure increases the complexity of pre-processing, whereby unnecessary characters must be removed or carefully replaced. Moreover, the complexity of Arabic-language morphology tends to increase the set of features, syntax, and semantic structures, which is not germane to authorship analysis tasks. Therefore, although the language provides the flexibility of numerous features, most of them are either not used or are not enough for related tasks. One implication of this is the limited number of Arabic-language authorship identification studies and the minimal number of datasets that are available for the Arabic language (Alqahtani & Dohler, 2022b).

3 Methodology

The most recent work on Arabic AV using linguistic features (stylometric features and TF-IDF) gave promising results (Alqahtani & Dohler, 2022a), and this study seeks to build on this by investigating the effect of using other features that help to extract more tweets, and understand the context of tweets without being word-dependent. In this work, we will continue the work on the same dataset and use the same stylometric features to find the effect of the investigated features on verifying authorship.

The concept is creating a table of words (each in different rows) that carry specific values for each column/feature. The aim is to explain the words' meanings and identify their range of closeness or divergence from other words. We created an Arabic knowledge-base in the form of a large table, whereby each word in a row is described with a set of features (columns), each of which carries a number between 0 (if the column is the complete opposite of the word) and 1 (if the column is the exact meaning of the word). Other values are explained in more detail in section 3.2.

As stylometric features are considered to be essential features for AV tasks, we extend the work in (Alqahtani & Dohler, 2022a) in addition to using our novel AraKB model. The results give an indication about the effect of using this technique to verify authorship with special relevance to very brief texts (specifically tweets).

3.1 Dataset

As a part of our investigation, and in order to have comparable results from our experiments, we used the same dataset as in our previous work (Alqahtani & Dohler, 2022a). The dataset contains 100 Twitter users, tweeting in a mixture of Gulf dialect and MSA. The total number of tweets in the corpus is 375,428, with a maximum of 3000 and minimum of 1000 tweets per user. For the knowledge-based model, as this experiment is fully accredited on words, the first step was to prepare the words included in the knowledge-based table. Arabic language words for classical and MSA alone are counted in the millions (Jalaluddin Al-Suyuti, 1998), in addition to newly generated words in various forms of colloquial Arabic. Creating one table containing all Arabic words is impossible, thus we extracted the 1000 most-used words in the dataset to employ them in the experiment.

It is important to note that when extracting the most used words in the dataset we found many Arabic stop words; while used heavily in writing, stop words usually have a negligible impact on the meaning of sentences (Bouzoubaa, Baidouri, Loukili, & Yazidi, 2009). Although stop words are usually eliminated in the data processing phase in some NLP tasks, such as information retrieval, in order to reduce noise, we argue that some stop words some words are important to

keep, and always make a difference in the sentence, particularly negation words.

Negation words play vital role in changing the sentence meaning completely. For example, the word (not/ليس) when added to any word will give the opposite or negation of it, and therefore change the whole meaning. Hence, we kept all the negation words, such as (لم, لا ليس) and any dialectical word that carries a negation meaning, such as (مو ,ماني , مافي).

3.2 AraKB creation

After the words were extracted and the unimportant stop words were eliminated, the table was created for the top 1000 most-used words in the dataset. However, we found some words that could not be included, such as names, usernames, and English words, which were consequently ignored.

It is important to point out that we treated emojis and punctuation the same as words, because in this experiment we aim to not only determine their existence but also investigate the meaning that they carry. In tweet communication, such features assume a particular potency and relevance to particular authors' styles and sentiments, which can often be expressed more fully by an emoji or a particular punctuation mark. The total number of the actual words was 895 words, 15 punctuations, and 90 emojis.

Regarding the number of features (columns), we tried to create as many features as possible to describe the words in a detailed way. An insufficient number of descriptions would be insufficient to enable the model to predict meaning, while a greater number of describing words conversely yields more accurate results.

In this experiment, we created a big table which includes 1000 rows and 100 columns, wherein each row carries one word, and each column contains one feature. These 100 features were about common status, words, or adjectives that would describe the meaning of different words. When writing the features, the following issues were considered:

1. The features did not contain any word and its opposite, to avoid repetition. For example, we did not need to have two features **(Cold and Hot)**, because when we give the feature **Cold** the value 0, that would give the same meaning of the word **Hot**.
2. We wrote the feature names in English language in order to make the work readable and understandable by non-Arabic readers.

Each word is represented by a value that distinguishes it from other words. Each word in a row is described with a set of features (columns), each of which carries a number between 0 (if the column is the complete opposite of the word) and 1 (if the column is the exact meaning of the word). In addition, features that do not take the values 0 or 1 will take floating point numbers (between 0-1), based on the relatedness of the feature with the word. For features that are not applicable for the word, or which do not carry a yes/no answer, the value "NA" (not applicable) was assigned.

Features

Words	POS	FW	Negation	Emoji	Puncutation	Abbreviation	Word correctness	Formal
شكرا	Interjection	1	NA	0	NA	NA	0.4	0.5
الحب	N	0	NA	0	NA	NA	NA	0.5
الملك	N	0	NA	0	NA	NA	NA	0.5
كفووو	NA	0	NA	0	NA	NA	NA	0.4
كورونا	N	0	NA	0	NA	NA	NA	0.5

Figure 1: Sample of the AraKB data.

The purpose is to enable the model to recognize the approximate meaning of the word by having more description and the meaning of each word, rather than merely acknowledging the existence of the word itself. Figure 1 presents examples of the words each and their description (features). This method allows understanding the sentiment features and semantic relationships of the context, which might help to understand the user's pattern. Through the combination of the features' values in each tweet, the model predicted the context and the user's style of writing.

It is important to note that the process of entering this voluminous information was not random, but followed a specific method, as discussed in the next section.

3.3 AraKB annotation

This experiment aims to understand the meaning of words among Gulf Arabic speakers. Consequently, the AraKB table was created using words taken from the collected tweets and was compiled manually, to produce data that mimics real language on social media. The table was filled by the researcher and reviewed by an Arabic linguist to ensure an accurate description of the words, and that it was a reliable and realistic reference for the use of Gulf Arabic dialect words.

The linguist is an Arabic teacher who holds a Master's degree in Arabic language and literature. In addition, the linguist is active on different social media accounts (including Twitter), and is therefore familiar with the use of words and synonyms among Twitter users. More importantly, the linguist speaks the Gulf dialect (same dialect of the dataset), to ensure that they understand and describe the words and their full semantic and contextual implications as intended by Twitter users.

Regarding the features, a list was created to describe words from different aspects. For the sake of better explanation about the nature of features, they were divided into three categories, as explained in Appendix A.

3.4 Evaluation

During the creation of AraKB and annotating the words, some agreements and disagreements emerged between the researcher and the linguist on the values/description of the words. All the words themselves were kept as they are, and each produced a different table of values. Consequently, a method was needed to assess the level of agreement between the annotators in order to evaluate the quality of the data. As the data comprises nominal values, Cohen's kappa coefficient (for inter-rater reliability) was applied to measure the reliability of AraKB data.[2]

We calculated the number of complete agreements for all values. For each word, we counted how many features are identical in values between the annotators (agreement values). Any different values filled by the annotators were considered to be disagreements, regardless of the difference between values, such as different floating point numbers. For example, a field could carry the value 0.3 for Annotator #1, and 0.7 for Annotator #2. In that case, any difference was considered as a disagreement, and the same applied on all fields/values. Kappa is measured through the following equation:

$$k = \frac{p_0 - p_e}{1 - p_e} \quad (1)$$

Where p_o is the actual values of agreement among the annotators divided by the total number of values, and p_e it is probability of agreement between columnists, calculated as follows:

$$p_e = \sum_q \frac{n_{A1q}}{i} \times \frac{n_{A2q}}{i} = \frac{1}{i^2} \sum_q n_{A1q} \times n_{A2q} \quad (2)$$

After the value of k is calculated, the result is categorized to a specific level of agreement (McHugh, 2012), which is shown in Table 1

Cohen's kappa statistics	Level of agreement
≤ 0	No agreement
0.1 – 0.20	Poor agreement
0.21 – 0.40	Fair agreement
0.41 – 0.60	Moderate agreement
0.61 – 0.80	Substantial agreement
0.81 – 0.99	Almost perfect agreement
1	Perfect agreement

Table 1: Interpretation of Cohen's Kappa value.

The Kappa coefficient was measured for 100,000 values. In the case of our data, due to the large number of values, the level of agreement between the annotators was measured based on each column (feature). However, it is important to note that there were some columns that had 100% agreement (for the language-constant features), which therefore cannot carry different values (as described in Appendix A). In addition, 34 features of the second category also had 100% agreement.

The different values of the features and possible disagreements happened in the third category (60 features), which had the possibility to carry different values based on dialect, context, or different opinions of the annotators. In order to give more realistic and interpretable values, scales were agreed for some features to measure the relatedness of the word to the specific feature. For example, the feature Dangerous had a range of values based on how "dangerous" the word is, thus the value 1 was given for words like **Drugs, Kill**, and very dangerous things that could cause death. Values of 0.9-0.5 were assigned for other dangerous things that would not necessarily cause death, such as the words **Disease** and **Scorpion**. Values of 0.4-0.1 were given for things that may cause death if used wrongly, like **Car, Technology,** etc. Lastly, the value 0 was given for very safe things, such as the word **Shirt**.

The same process was applied for most of the features. It is important to note that there were

[2] Introduced by Jacob Cohen in 1960.

some words that carry a different meaning among Saudi users. For example, Thursday/الخميس is the beginning of the weekend at Saudi Arabia and most Gulf countries, so this word is usually used in semantic clouds connoting the status of fun and partying. Consequently, the word took the value 1 for the feature Interesting, based on its use in our data. Another example is for words used in informal contexts in different ways, such as the word بيض/egg. It is axiomatic that this word should be described as Food, but after an observation of its usage among the Saudi social media users, it was clearly used to express a state of boredom or something that is not interesting, thus it was given the value 0 for the feature Interesting.

After both annotators filled all fields of AraKB separately, we applied Kappa statistics only on the features of the third category (60 features), whose values could carry agreement and disagreement. For that a number of 60,000 features (60 features × 1000 words) were calculated by Cohen's kappa, which resulted in the findings reported in Table 2.

Total of fields	Agreement	Disagreement	Kappa Value
60,000	59,925	75	0.99

Table 2: Cohen's Kappa and inter-rater agreement.

Kappa value is 0.99, which is considered to be almost perfect based on the interpretation of Cohen's Kappa value, which gives the table more reliability, whereby it can be used in other works related to Gulf Arabic texts. This value was achieved as there is a specific measurement for each field, as explained earlier.

However, it is important to state that most of the fields had the NA value, because not all the features are applicable to describe the words, which explains the low number of disagreement values between the annotators. A total of 2,818 exact values were filled with numbers other than NA.

4 Experiments

4.1 Experimental setup

Unlike the use of TF-IDF features in the work of (Alqahtani & Dohler, 2022a), which depended on the existence of the words, the main purpose of AraKB is that the model will recognize the approximate meaning of the word by the set of features. Therefore, it can verify the users through the repeated status, emotions, and expressions reflected in their written texts.

Firstly, we needed to convert our AraKB Excel table into a form that would be usable in our code. A huge dictionary was created that contains all 1,000 words as keys (keys#1), whereby each word has its own smaller dictionary that has the set of features as keys (keys#2), each with their values (0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1, or NA). The dictionary was created for all columns except the POS column, as it has different values than the other columns (i.e., verb, noun, adjective, etc.), which was annotated manually and will be explained separately.

After that, we calculated the vectors by dividing the tweet into single words (tokens). A function was created that takes each word of the tweet and finds if it exists in one of the (keys #1). If the word exists, its dictionary with the key#2 is recalled, and their values are added. This process is repeated for every word in the tweet, then the values of all words in the tweet are calculated by taking the average of each feature's value of the words. If no word in the text exists in the table, the model will ignore the text.

We cannot take the average of the categorical values of the feature/column (POS), as this column has a different part of speech tags that represent words such as nouns, verbs, and adverbs, etc. Each value of the POS column has a separate column, so that each word will have a value in the related POS column (as previously set in the AraKB).

Another function was created related to the Negation feature, which contains the most used negation words that exist in the dataset (either MSA or Gulf dialect). The purpose of this feature is to reverse the meaning of the word following the negation. After calculating the vectors, the value of word coming after negation will be the opposite and will be with minus. For example, the word "صادق/honest" has the value 1 in the Honesty feature, but if it comes after the word "ليس/not" then the value for Honesty will be converted to -1. Lastly, the values of the Honesty feature will be averaged with other words existent in the tweet, therefore the level of "Honesty" will be reduced in the whole tweet.

To sum up, the values of each word of the tweet were extracted and averaged with other values. Figure 2 shows the concept of the calculation. The model scanned the tweet and found the words (المدرسة/المتعة/الطفل) that exist in our AraKB, thus it took each value for the column 1 (feature 1) and calculated the average values for all existent words, which gave the value 0.53. The same was repeated for all columns, until a list of vectors representing each tweet was compiled.

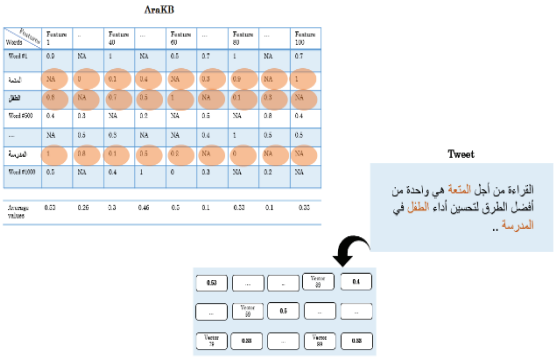

Figure 2: Process of converting the tweets to vectors.

Using this method might enable the model to ascertain the extent to which the tweet carries values of each feature written by each user, whereby it might be able to identify a pattern about how the user expresses their thoughts/emotions in their writing. In this experiment we used a previously tested stylometric feature (Alqahtani & Dohler, 2022a) in addition to our novel AraKB features.

4.2 Experimental results

As stated earlier, this experiment was based on the dataset used by (Alqahtani & Dohler, 2022a), and with same setting used in the previous experiments (train/test ration, the used classifier, CV 5-folds, etc.), in addition to conducting the same pre-processing steps in order to have comparable results.

Using the best performance algorithm (Gradient Boosting), our experiment showed a 2% improvement in performance (accuracy and F1-score) when adding the AraKB features to the previous stylometric features, as opposed to using the latter alone. Table 3 and Table 4 compare the results of using the stylometric features in the previous study (Alqahtani & Dohler, 2022a) and the results of the same tested dataset when using the AraKB features (respectively)

Feature	Avg F1	Avg recall	Avg precision	Avg accuracy
Stylometric	0.75	0.76	0.75	0.75
Stylometric + TF-IDF	0.77	0.75	0.79	0.77

Table 3: Results of using stylometric features by (Alqahtani & Dohler, 2022a).

Feature	Avg F1	Avg recall	Avg precision	Avg accuracy
Stylometric + AraKB	0.77	0.77	0.77	0.77

Table 4: Results of using AraKB and Stylometric features.

5 Discussion

Looking at the results of using a combination of stylometric and AraKB features shows a similarity in average results with the results of previous experiments that used a combination of stylometric and TF-IDF features. This indicates that using AraKB gives similar performance to using the TF-IDF features.

However, it is important to note that AraKB features contained only 1,000 Arabic words, due to the laborious and time-consuming individual efforts entailed. It is assumed that the outcomes would be substantially improved by adding many more words that an author would possibly write with. The purpose of this experiment is preliminary testing using AraKB features, to determine if these features enhance the performance of verifying authorship in short texts like Twitter posts.

One could argue that using AraKB is over-fitted to the dataset from which the list of words was derived. However, we have selected the thousand most prolifically used words of the whole dataset, which is actually far from being a reason of overfitting. This is because choosing the most repeated words entails that most of the users have used these words, therefore these words are not considered to be user-distinctive (i.e., they reflect

homogenous use of language). In addition, our approach considers the meaning of words averaged with others in the same tweets, which means that it is not word-dependent like other content-dependent features, such as TF-IDF or BOW.

6 Conclusion

The limited work on Arabic AV texts shows the need to investigate more features that could enhance the verification process. In this experiment, we prove that creating features that represent the word's meaning, as in AraKB, does help to effectively verify the authorship, and might be helpful in other NLP tasks, such as sentiment analysis.

Future work on AraKB might extend the number of Arabic words, which will definitely improve its performance. In addition, further studies should investigate the influence of each of AraKB features, as we could focus only on the most influential ones.

Acknowledgments

The authors of this paper would like to thank the linguist Maha Alqahtani for taking the time and effort to manually review AraKB.

References

Ahmed Abbasi and Hsinchun Chen. 2005. Applying authorship analysis to extremist-group Web forum messages. *IEEE Intelligent Systems*, *20*(5), 67–75. https://doi.org/10.1109/MIS.2005.81

Bouzoubaa, K., Baidouri, H., Loukili, T., & Yazidi, T. El. (2009). Arabic stop words: Towards a generalisation and standardisation. In *Knowledge Management and Innovation in Advancing Economies: Analyses and Solutions - Proceedings of the 13th International Business Information Management Association Conference, IBIMA 2009*, *3*, 1844–1848.

El-Said Badawi. 1996. Understanding Arabic: essays in contemporary Arabic linguistics in honor of El-Said Badawi. American Univ in Cairo Press.

Fatimah Alqahtani and Mischa Dohler. 2022a. Investigating Predictive Features for Authorship Verification of Arabic Tweets. IJCSNS, 22(6), 115.

Fatimah Alqahtani and Mischa Dohler. 2022b. Survey of Authorship Identification Tasks on Arabic Texts. In *ACM Transactions on Asian and Low-Resource Language Information Processing*.

Hosein Azarbonyad, Mostafa Dehghani, Maarten Marx, and Jaap Kamps. 2015. Time-Aware Authorship Attribution for Short Text Streams. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages. 727-730.

Hossam Ahmed. 2017. Dynamic Similarity Threshold in Authorship Verification: Evidence from Classical Arabic. In *Procedia of The 3rd International Conference on Arabic Computational Linguistics*, pages 145–152. https://doi.org/10.1016/j.procs.2017.10.103

Hossam Ahmed. 2018. The Role of Linguistic Feature Categories in Authorship Verification. In *Procedia of The 4th International Conference on Arabic Computational Linguistics*, *142*, pages 214–221.

Hossam Ahmed. 2019a. Distance-Based Authorship Verification Across Modern Standard Arabic Genres. In *of the 3rd workshop on arabic corpus linguistics,* pages 89-96

Hossam Ahmed. 2019b. Sample Size in Arabic Authorship Verification. In *Proceedings of the 3rd International Conference on Natural Language and Speech Processing,* pages 84-91. Association for Computational Linguistics.

Jalaluddin Al-Suyuti. 1998. Al-Mazhar fi 'ulum al-lughat wa 'iwa'aha. Almaktabat aleasria.

Karin C. Ryding. 2005. A Reference Grammar of Modern Standard Arabic. Cambridge university press. https://doi.org/https://doi.org/10.1017/CBO9780511486975

Khaled Shaalan and Hafsa Raza. 2009. NERA: Named entity recognition for Arabic. In *Journal of the American Society for Information Science and Technology*, *60*(8), pages 1652–1663. https://doi.org/10.1002/asi.21090

Kim Luyckx and Walter Daelemans. 2011. The effect of author set size and data size in authorship attribution. *26*(1), pages 35-55 https://doi.org/10.1093/llc/fqq013

Leah S. Larkey and Margaret E. Connell. 2001. Arabic Information Retrieval at UMass in TREC-10. In *The Tenth Text REtrieval Conference*.

Mary L. McHugh. 2012. Interrater reliability : the kappa statistic. In *Biochemica Medica*, *22*(3), pages 276–282. Retrieved from https://hrcak.srce.hr/89395

Mohammad Al-Sarem, Walid Cherif, Ahmed Abdel Wahab, Abdel-Hamid Emara, and Mohamed Kissi. 2018. Combination of stylo-based features and frequency-based features for

identifying the author of short Arabic text. In *of the 12th International Conference on Intelligent Systems: Theories and Applications,* pages 1-6. https://doi.org/10.1145/3289402.3289500

Sushil Kumar and Mousmi A. Chaurasia. 2012. Assessment on Stylometry for Multilingual Manuscript. In *IOSR Journal of Engineering*, *2*(9), pages 1–6. https://doi.org/10.9790/3021-02910106

A Appendices

Features specifications:

1. Language-constant features: Where we had six features that are considered to be constant in the language, which are known and cannot be considered in terms of personal opinion. These features are: Part of Speech, Function word, Negation, and Punctuation. Filling these words was based on previous knowledge of Arabic grammar. In addition, the features (Abbreviation and Emoji) took either the value 0 or 1 value based on the existence of the feature in the word. The following provide more details and examples about each feature and explanation of why they were considered as constant features.

2. Features that carry a yes/no answer. There were 34 features that only took values of 0, 1, or NA; the values cannot be a number in between 0-1. These features are: (Human, Alive, Female, Animal, Body-part, Food-related, Time-related, Place-related, Eaten, Past, Work/study, Question, Nationality, Weather, Prayer, Compare, Place, Time, Number, Many, Media-content, Listing, Quoting, Calling, Sport, Art, Policy, Literature, Religion, Science, Travel, Economy, Law, and Technology). These features are either applicable on the word or not, so they would take either the value 1 or 0.

3. The other features. The remaining 60 features may carry range of different values (e.g., 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1, or NA) based on the word's relatedness to the features. These features are: (Formal, Word correctness, Long, Strong, Expensive, Dangerous, Ability, Shyness, Like, Peaceful, Loyalty, Excellence, Privacy, Necessity, Reality, Desire, Request, Rich, Big, Normal, Beautiful, Smart, Useful, Cause-death, Noisy, Cold, Heavy, Thankful, Youthful, Romantic, Agree, Happy, Angry, Welcome, Sarcastic ,Similar, Scary, Disgusting, Well-known, Crime, Childish, Optimism, Simple, Comfortable, Interesting, Healthy, Surprised, Wonder, Argue, Certainty, Emphasis, Honesty, Emotional , Laugh, Abusive, Racist, Shame, Wish, and Royal).