

Arabic authorship attribution on Twitter: what is really matters?

Anoual El Kah¹, Aymane El Airej², Imad Zeroual³

¹Department of Computer Science, Faculty of Sciences, Mohamed First University, Oujda, Morocco

²Department of Computer Science, Moulay Ismail University of Meknes, Meknes, Morocco

³Department of Computer Science, Faculty of Sciences and Techniques, Moulay Ismail University of Meknes, Meknes, Morocco

Article Info

Article history:

Received Jul 2, 2022

Revised Aug 30, 2022

Accepted Sep 16, 2022

Keywords:

Arabic tweets

Authorship attribution

Bag-of-n-grams

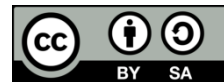
Feature extraction

Stylometric features

ABSTRACT

Recently, authorship attribution (AA) of online social networks texts has gained more attention. However, since 2015, when the first work that addressed the AA of Arabic tweets was published, we found that nothing much has been done after that. Thus, the current paper presents an extensive study that investigates the effects of various factors on the AA of Arabic short-texts, especially tweets. This led to a proposed architecture in which the AA accuracy is examined depending on the size of the training dataset, the number of classes covered, the text processing techniques applied, the methods used for both feature selection and extraction, and finally, the classifier implemented. As a result, we performed 792 different tests. The highest accuracy recorded is 97.4%, and it is among the best results published so far.

This is an open access article under the [CC BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



Corresponding Author:

Anoual El Kah

Department of Computer Science, Faculty of Sciences, Mohamed First University

Oujda 60000, Morocco

Email: elkah.anoual.mri@gmail.com

1. INTRODUCTION

Authorship attribution (AA) is a type of classification problems which considers a set of authors as classes. AA aims to accurately attribute a disputed or unknown text to its. This task is deeply rooted in history since identifying the author of ancient texts has always been the center of attention of linguists [1]. During the last decade, most researchers have addressed text in online social networks, primarily for sentiment analysis [2]-[4]. However, AA of social networks texts has gained more attention bringing new trends to this field.

With more than 10.5 million tweets per day [5], Arabic is among the top five dominant languages on Twitter [6]. However, the first work dealing with AA of Arabic tweets is attributed to a publication dating back to 2015 [7]. In fact, during the following three years, barely ten published works that addressed Arabic AA [8]. Unfortunately, till writing this manuscript, we found that nothing considerably had changed. As for many languages, Arabic AA is understudied in light of artificial intelligence and still not well-investigated [9].

This paper investigates the effects of various factors on the AA of Arabic short texts, especially tweets. The first investigated factor is the data size or, more precisely, the number of tweets attributed to each author in the training dataset. Followed by the number of authors (i.e., classes) covered in the used dataset. Further, we investigated the influence of the features selected to represent the text. In this regard, we first examined the original words used as the baseline; then, these words were replaced by linguistic and stylometric features. The linguistic features used are either morphological such as stems and lemmas or syntactic such as the part-of-speech (PoS) tagset. The next factor addressed is the feature extraction methods

implemented such as the term frequency-inverse document frequency (TF-IDF) and Countvectorizer. The final factor discussed is the algorithm selected to build the classification model. To This end, we investigated the performance of the state-of-the-art classification models namely support vector machines (SVM), random forests (RF), Naïve Bayesian (NB), and their combination.

The five coming sections contain a detailed description of our investigation. The second section introduces related works. In the third section, we describe in detail our methodology adopted to perform the current study. The fourth section presents the results recorded for each mentioned factor. In the fifth section, we discuss and illustrate our findings. Finally, we draw the conclusion and perspectives in the sixth section.

2. RELATED WORKS

In 2015, Albadarneh *et al.* [7] performed the earlier work that handled the AA of Arabic tweets. First, they build a dataset that comprises 53,205 tweets posted by 20 different authors. Most of these tweets are written in dialectal Arabic. Then, the feature extraction process was performed using the TF-IDF method. Those features were next inserted into the hadoop distributed file system. They finally used the NB-based classifier to identify the authors of the anonymous tweets, and the best accuracy recorded was 61.6%.

A second study [10] has fetched 37,445 tweets from 12 famous Arabic authors on Twitter. The study uses a combination of uni-gram, stylometric, and linguistic features. Subsequently, three well-known classifiers, namely decision tree (DT), SVM, and NB, were applied independently. The best performance, i.e., 68.67%, was obtained by the SVM-based classifier using all the features combined. Later, the study has been extended [11] by applying new feature selection techniques like SubEval, principal component analysis (PCA), ReliefEval, CorrEval, and InfoG. However, the performance recorded by the classifier was slightly better (68.90%). In 2018, an extensive study [8] investigated the performance of four classifiers, namely RF, SVM, DT, and NB, using n-gram and stylometric features under several conditions. The main findings were that applying n-grams technique leads to better results, while the best accuracy rate, i.e., 94%, was obtained by the RF classifier.

A recent study [12] attributed the authors of offensive and inappropriate Arabic tweets. The authors compiled 20,357 tweets of 134 users from different Arab countries. The tweets were compiled from users who posted their disappointment with the Nicki Minaj’s performance that was supposed to be held in July 2019 in Saudi Arabia. Then, their classifier implemented a document clustering based on document index graph (DIG) model and PCA as a feature selection method. The accuracy rate reported is around 83%.

The last example will be a study [13] that aimed to benefit from using a bagging model to improve the accuracy of Arabic AA on Twitter. The authors compared the performance of their bagging classifier with three single learners, namely NB, SVM, and DT. As a result, the bagging classifier outperformed the other single classifiers by obtaining the best performance (95,03%). It is worth mentioning that ensemble methods have also shown effectiveness when applied for AA of fatwas written in Arabic [14].

3. METHOD

Our overall methodology consists of five main stages. The first stage is preparing four sub-datasets. The second includes various text processing techniques. Next, we identify the features representing the original text using stylometric and linguistics features. In the fourth stage, only relevant features are kept using two different methods, TF-IDF and Countvectorizer. These top-ranked features will be fitted to the classifiers in the last stage. Figure 1 displays our overall methodology.

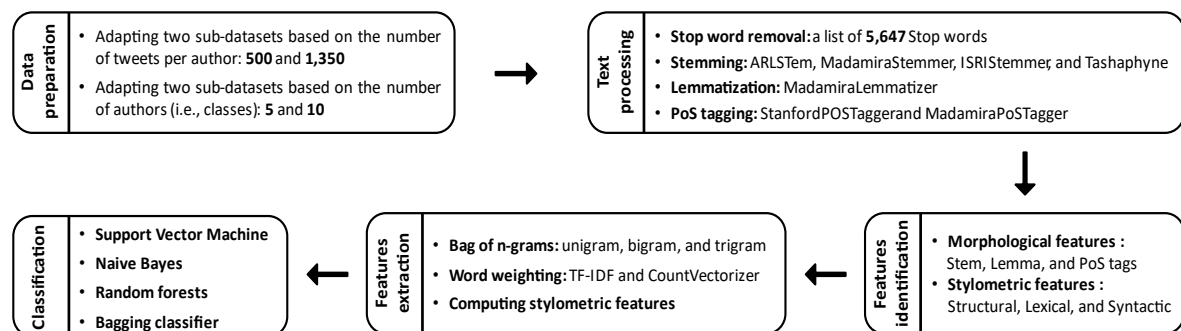


Figure 1. Overall methodology for the proposed study

3.1. Dataset

Our dataset is part of the BZU-ECE corpus [15]. This latter is a collection of 71,391 Arabic tweets posted by 44 different authors. We prepared four sub-datasets to investigate the effect of increasing training set size and the number of classes on the performance of AA classifiers. As results, the first sub-dataset includes 10 authors and 1,350 tweets per each. The second includes 10 authors and 500 tweets per each. The third includes 5 authors and 1,350 tweets per each. Finally, the fourth includes 5 authors and 500 tweets per each.

3.2. Text processing techniques

The effect of the following text processing techniques on the AA of Arabic tweets is investigated:

- a) Stop words removal: Although stop words are not related to a text's subject, their use may be related to the author's writing style. Therefore, we investigate this hypothesis by studying the effect of a stop words removal task on the performance of AA classifiers for Arabic short texts.
- b) Stemming: We used four different Arabic stemmers that were previously under investigation for their benefits for Arabic text classification. These stemmers are ARLSTem v1.0 [16], Tashaphyne, integrated system of rice intensification (ISRI) stemmer [17], and the stemmer included in Madamira [18].
- c) Lemmatization: Lemmatization has recently proved to be beneficial for Arabic text classifiers [19]-[21]. However, Lemmatization's impact is rarely investigated for Arabic AA. Hence, we included in this investigation using Madamira's lemmatizer.
- d) Part of speech tagging: Very few works have involved the PoS tagging in Arabic text classification in general [22], and in Arabic AA in particular [8], [10]. Thus, two robust Arabic PoS taggers have been used, namely the Stanford's PoS tagger [23] and the PoS tagger of Madamira.

3.3. Stylometric features

Stylometric features are intensively used for AA to quantify the writing style. In this study, we identified 25 features that represent three aspects of stylometric features namely, lexical, syntactic, and structural. Table 1 exhibits the 25 stylometric features adopted in this study.

Table 1. Stylometric features extracted

Category	Lexical and Character	Syntactic	Structural
Features	1. The number of characters	9. The number of Nouns	
	2. The number of words	10. The number of Verbs	
	3. The number of unique words	11. The number of Proper Nouns	
	4. The maximum length of the words	12. The number of Adjectives	
	5. The minimum length of the words	13. The number of Adverbs	
	6. The number of punctuations	14. The number of Pronouns	
	7. The number of digital numbers	15. The number of Prepositions	24. The average word length
	8. The number of Foreign (non-Arabic) words	16. The number of Feminine	25. The average sentence length
		17. The number of Masculine	
		18. The Number of Singular Words	
		19. The Number of Plural Words	
		20. The Number of Dual Words	
		21. The Number of 1 st Person	
		22. The Number of 2 nd Person	
		23. The Number of 3 rd Person	

3.4. N-grams models

N-grams or bag-of-n-grams is a language model that uses statistical techniques to learn the probability distribution of tokens (words). It is considered as an extension of bag-of-words in which the sequence of n tokens is calculated. Surveying the literature, the most n-grams sequences used are 1-gram, 2-gram, and 3-gram. To make our study more extensive, we used all these three types of n-grams models.

3.5. Feature extraction

For text classification, feature extraction is converting primary textual content into numerical features that will be handled by the classifier. In the current study, we investigate the effects of using both Countvectorizer and TF-IDF as feature extraction methods. Countvectorizer and TF-IDF are commonly used as feature extraction methods. Countvectorizer counts the frequency of each word in the processed data; whereas, the TF-IDF measure is calculated by multiplying the occurrence probability of a term in a single tweet (i.e., TF) and the inverse log of the number of tweets containing that term (i.e., IDF). Using such statistical methods have shown to be beneficial for AA and outperformed the state-of-the-art stylometric features [24].

3.6. Classification algorithms

In this stage, we used three cutting-edge classifiers, namely NB, SVM, and RF to evaluate the effects of the aforementioned factors. These classifiers are first implemented individually. Then, they are used to build a Bagging classifier. This Bagging classifier compound the ensemble predictions generated by the three classifiers to get the final prediction based on a majority voting procedure. In another words, it selects for each anonymous tweet the most voted author. If the given authors from the three classifiers are unlike; then, we select the author given by the most accurate classifier that outperformed the others under the same conditions.

4. EXPERIMENTS AND RESULTS

In this study, we investigated the effects of six main factors on the AA task under several conditions. This results a total number of 792 tests, and we recorded the accuracy rate of each test. Retrieving insights from these many rates is complicated; therefore, we represent the findings in brief and comprehensible forms in the following sub-sections.

4.1. Effect of the training data size

The first factor to investigate is the size of the training dataset. We prepared two datasets that comprises 1,350 and 500 tweets. All the 792 classifications are covered in this comparison. Then, we simply compared the two accuracies achieved by each classifier under the same conditions except for the training data size used to train the classifier. Thus, in 308 cases (78%), the classifiers achieved the best accuracy when 1,350 tweets are used; whereas, only in 88 (22%) cases, the classifiers performed better when 500 tweets are used.

4.2. Effect of the classes' number

The same as in the previous comparison, we have 396 comparison cases. Between the performance of each classifier using 5 classes and using 10 classes while fixing all the other factors. In the end, we found that all classifiers in all cases (100%) achieved their highest accuracy using only 5 classes.

4.3. Effect of text processing techniques

Selecting the appropriate linguistic feature that can represent the original text is still intriguing researchers working on Arabic text classification [19], [20]. Therefore, we used 10 different linguistic features to represent the tweets' text. The baseline represents the original words as they are in the tweets. Further, the same baseline words are used but after removing stop words. Then, we replaced the tweets' words by their stems using various Arabic stemmers namely ISRI, Tashaphyne, ARLSTem, and Madamira stemmer. Also, we used the lemmas generated by Madamira. The PoS tags are also involved using two tagsets, those generated by the Stanford PoS tagger and those of Madamira. Finally, the combination of all these features is used.

Each of these 10 linguistic features is used in classifications that involve two sub-datasets, two groups of classes, three classifiers, three bag-of-n-grams, and two feature extraction methods. Consequently, $10 \times (2 \times 2 \times 3 \times 3 \times 2) = 720$ accuracy rates are reported. However, we calculated the averages of these accuracies to make these results easy to understand for the reader. Thus, Table 2 presents those results for each classifier.

Table 2. Average and highest accuracy achieved by the classifiers using linguistic features

Features selected	Classifiers	Average accuracy	Highest accuracy	Features selected	Classifiers	Average accuracy	Highest accuracy
Baseline	SVM	85.37	92.6	Madamira Stems	SVM	65.53	77.9
	NB	87.62	93.2		NB	66.35	75.33
	RF	79.86	92.6		RF	60.61	71.84
Baseline with stop words removal	SVM	86.37	92.6	Madamira Lemmas	SVM	64.98	78.18
	NB	87.74	93		NB	66.35	76.28
	RF	78.89	85.7		RF	60.60	71.57
ISRI Stems	SVM	83.42	92.2	Stanford PoS tagset	SVM	37.18	50.54
	NB	84.70	89.85		NB	31.19	40.6
	RF	76.96	85.18		RF	35.93	48.4
Tashaphyne Stems	SVM	84.36	92.2	Madamira PoS tagset	SVM	41.29	53.72
	NB	85.94	91		NB	34.01	44
	RF	77.68	86		RF	38.89	50.54
ARLSTem Stems	SVM	85.16	92	Combined Features	SVM	61.17	87.36
	NB	86.83	91.4		NB	81.09	89.45
	RF	78.24	86		RF	78.47	85.81

According to Table 2, the best average accuracies are achieved by all the classifiers when the baseline words are used as features. These average accuracies are slightly increased when stop words are removed (SVM and NB). On the contrary, they are decreased when the baseline words are replaced by their stems (similar to [25]), lemmas, PoS tags, or by their combination. Additionally, 8 out of the 10 best average accuracies are recorded by the NB classifier while only 2 have resulted from the SVM. However, this latter achieved 7 out of the 10 highest accuracies; whereas the remaining 3 are obtained by the NB classifier. Another observation is that the less accuracies are achieved when the stems or lemmas of Madamira are used. Regarding the PoS tagsets, the Stanford PoS tagger seems to lack behind Madamira.

4.4. Effect of stylometric features

The effects of the aforementioned three types of stylometric features and their combination are investigated using the same two sub-datasets, two groups of authors, and the three classifiers. This led to performing $4 \times (2 \times 2 \times 3) = 48$ different classifications. The findings are summarized in the following Table 3 that presents only the average accuracies alongside the highest accuracies that have resulted from all the classifiers.

As seen in the table, the best average accuracies are achieved by the SVM and the NB classifiers when only the syntactic features are used; whereas, the best average accuracy, which resulted from the RF classifier, is recorded when all the stylometric features are combined. Except this latter achievement recorded by the RF classifier, the SVM classifier was the best in all other cases.

Table 3. Average and highest accuracies achieved by the classifiers using stylometric features

Stylometric features	Classifiers	Average accuracy	Highest accuracy
Lexical features	SVM	38.97	49.11
	NB	32.51	40
	RF	34.98	44
Structural features	SVM	34.23	43.18
	NB	27.76	38.66
	RF	29.50	38.51
Syntactic features	SVM	40.35	48.8
	NB	32.60	40.4
	RF	37.60	48.6
Combined features	SVM	38.38	48.9
	NB	25.23	33.81
	RF	39.28	48.63

4.5. Effect of n-grams

To investigate the effect of bag-of-n-grams on the AA classifiers, we have adopted the uni-gram, bi-grams, and tri-grams to learn the probability distribution of tokens. This investigation is based on the same 720 accuracy scores mentioned previously while the linguistic features were used. Similarly, we calculated the average accuracies that have resulted from the classifiers for each language model of these n-grams. Table 4 summarizes the results achieved by each classifier using those n-grams. According to Table 4, the best average accuracy (71,96%) and the highest accuracy (93,2%) are reported when the tri-grams model is used. Furthermore, the NB classifier outperforms the others in terms of either the best average accuracies or the highest accuracies achieved.

Table 4. Average and highest accuracies achieved by the classifiers using bag-of-n-grams

N-grams	Classifiers	Average accuracy	Highest accuracy
Uni-gram	SVM	68.65	91.8
	NB	69.72	91.4
	RF	65.34	91.2
Bi-grams	SVM	70.02	92.6
	NB	71.86	92.6
	RF	67.18	92.4
Tri-grams	SVM	69.78	92.6
	NB	71.96	93.2
	RF	67.32	92.6

4.6. Effect of feature extraction methods

Among the factors investigated in this study is the method implemented for the feature extraction stage. As noted in section 3.5, we have adopted the Countvectorizer and the TF-IDF. Both methods were

used in all the classifications performed. Likewise, we calculated and listed in Table 5 the average as well as the highest accuracies that have resulted from the classifiers for each feature extraction method.

According to Table 5, the average accuracies that have resulted from each classifier while using Countvectorizer or TF-IDF are quite similar. However, the best result is reached while using the Countvectorizer as a feature extraction method. Although the NB classifier achieved the highest accuracy, it also scored the lowest accuracy when implementing the TF-IDF method.

Table 5. Average and highest accuracies of the classifiers according to the feature extraction methods used

Extraction methods	Classifiers	Average accuracy	Highest accuracy
Countvectorizer	SVM	71.07	89.85
	NB	72.74	93.2
	RF	66.97	86.96
TF-IDF	SVM	70.19	92.6
	NB	72.04	92.4
	RF	67.47	92.6

4.7. Effect of classification algorithms

We implemented three classifiers namely NB, SVM, and RF separately under various conditions, which produced 240 accuracy results for each classifier. First, the classifiers were ranked according to their performance. Then, the output of the most accurate classifier is selected in the Bagging classifier when the given authors from the three classifiers are unlike. Consequently, the NB classifier was ranked first in 115 cases (48%), the SVM classifier in 106 cases (44%), and the RF classifier in only 19 cases (8%). Next, we calculated the average and the highest accuracies that have resulted from the classifiers separately as well as the results achieved by the bagging classifier. Table 6 exhibits the findings.

The full experiments showed that the performance of both NB and the SVM is generally quite similar in terms of the number of cases they ranked first and the average and the highest accuracy recorded. However, the NB classifier ranked first in the overall investigations when the three classifiers are implemented individually. Still, the best average (90,91%) and the highest (97,4%) accuracies reported in the whole investigation are those achieved by the bagging classifier. For completeness's sake, we provide a comparison of our best result with similar works. Table 7 exhibits the relative conditions of each study that led to its best accuracy rate. According to this comparison, the current study achieved the highest accuracy (i.e., 97.4%).

Table 6. Average and highest accuracies achieved by the classifiers

Classifiers	Average accuracy	Highest accuracy
SVM	69.48	92.6
NB	71.18	93.2
RF	66.61	92.6
Bagging	90.91	97.4

Table 7. Highest accuracies achieved by similar studies

Classifiers	Avg tweets per author	Nb. of authors	Features selected	N-grams	Feature extraction method	Classifier	Highest accuracy
[7]	2,660	20	Baseline	1-gram	TF-IDF	NB	61.6
[10]	3,120	12	Linguistics + Stylometric	1-gram	StringToWordVector	SVM	68.67
[11]	3,120	12	Linguistics + Stylometric	1-gram	SubEval, CorrEval, PCA, ReliefEval, and InfoG	SVM	68.9
[8]	747	2	Stylometric	1-gram	TF-IDF	RF	94
[12]	152	134	Stylometric	1-gram	PCA	DIG	83
[13]	1,479	45	Baseline	1-gram	TF-IDF	Bagging/SVM	95.03
Our model	500	5	Baseline	3-grams	Countvectorizer	Major voting (NB+SVM+RF)	97.4

5. DISCUSSION

Here is a quick recap of what is being done in the current investigation. This study investigated the influence of various factors on the entire AA procedure. Indeed, the results reported confirm that all six factors have an impact on the author's attribution efficiency. The list below summarizes the findings:

- a) Training data size: 78% of the experiments conducted led to improved accuracy when the number of tweets per author increased from 500 to 1,350.

- b) Number of authors: As the number of candidate authors increases, it leads to a decreased accuracy.
- c) Selected features: The investigation shows that the accuracy dropped for all the attribution performed when the baseline words were replaced by linguistic or stylometric features. Nevertheless, since the original dataset was already lemmatized and the stop words were removed, the influence of using the morphological (i.e., stems and lemma) and syntactic (i.e., PoS tagset) features on author's attribution may need further investigation using a different dataset.
- d) Bag-of-n-grams: 62% of the highest accuracies are recorded when the tri-grams model was used, 32% with bi-grams, and only 6% with the uni-gram model.
- e) Feature extraction: 60% of the highest accuracies are scored when the TF-IDF is used, whereas, 40% with the Countvectorizer. However, we should point out that both the best average and the highest accuracy in the entire investigation were recorded when the Countvectorizer method is implemented.

Classification algorithm: The NB and SVM classifiers shared the first and best ranks, while the performance of the RF classifier remains weak compared to them. However, the accuracy recorded by the bagging classifier in different experiments was always higher than those achieved by the other classifiers separately. After checking some samples, the reason may be because the three classifiers attribute different incorrect authors and these differences are exploited in combination to yield better results. Therefore, using the Bagging classifier is recommended, especially if involved classifiers generate different incorrect classes.

6. CONCLUSION AND PERSPECTIVES

This extensive investigation is among very few works that addressed the AA of Arabic tweets. Through this work, we studied the influence of various factors that are generally employed in text classification. The results showed that each factor has its impact on the entire classification procedure. To emphasize this claim, we conducted many comparative studies and the best accuracy rate of the author's identification was always recorded and discussed according to the factor under investigation.

This work is among the contributions that aim to fill the gap between Arabic AA and western languages. Despite the encouraging results obtained, more extensive and detailed investigations are still required. Besides, building freely available datasets is among our primary recommendations. Additionally, we highly recommend investigating what features are the best to represent the original text, especially for highly inflected languages like Arabic. Involving at least one feature extraction method seems to be sufficient to level up the performance. Finally, improved attribution is always still at hand by combining the outputs achieved by different algorithms using the Bagging classifier.




REFERENCES

- [1] S. Hriez and A. Awajan, "Authorship Identification for Arabic texts using logistic model tree classification," in *Science and Information Conference*, Jul. 2020, pp. 656–666, doi: 10.1007/978-3-030-52246-9_48.
- [2] E. T. Khalid, E. B. Talal, M. K. Faraj, and A. A. Yassin, "Sentiment analysis system for COVID-19 vaccinations using data of Twitter," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 26, no. 2, pp. 1156–1164, May 2022, doi: 10.11591/ijeecs.v26.i2.pp1156-1164.
- [3] P. Sugumaran and A. B. B. K. Uma, "Real-time twitter data analytics of mental illness in COVID-19: sentiment analysis using deep neural network," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 26, no. 1, pp. 560–567, April 2022, doi: 10.11591/ijeecs.v26.i1.pp560-567.
- [4] H. Elzayady, M. S. Mohamed, K. M. Badran, and G. I. Salama, "Detecting Arabic textual threats in social media using artificial intelligence: An overview," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 25, no. 3, pp. 1712–1722, Mar. 2022, doi: 10.11591/ijeecs.v25.i3.pp1712-1722.
- [5] S. O. Alhumoud, M. I. Altuwaijri, T. M. Albuhairei, and W. M. Alohaideb, "Survey on arabic sentiment analysis in twitter," *International Journal of Computer and Information Engineering*, vol. 9, no. 1, pp. 364–368, 2015, doi: 10.5281/zenodo.1099604.
- [6] T. Alshaabi, D. R. Dewhurst, J. R. Minot, M. V. Arnold, J. L. Adams, C. M. Danforth, and P. S. Dodds, "The growing amplification of social media: measuring temporal and social contagion dynamics for over 150 languages on Twitter for 2009–2020," *EPJ Data Sci.*, vol. 10, no. 1, Art. no. 1, Dec. 2021, doi: 10.1140/epjds/s13688-021-00271-0.
- [7] J. Albadarneh, B. Talafha, M. Al-Ayyoub, B. Zaqibeh, M. Al-Smadi, Y. Jararweh, and E. Benkhelifa, "Using big data analytics for authorship authentication of arabic tweets," in *2015 IEEE/ACM 8th International Conference on Utility and Cloud Computing (UCC)*, Dec. 2015, pp. 448–452, doi: 10.1109/UCC.2015.80.
- [8] M. H. Altakrori, F. Iqbal, B. C. Fung, S. H. Ding, and A. Tubashat, "Arabic authorship attribution: An extensive study on twitter posts," *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, vol. 18, no. 1, pp. 1–51, Nov. 2018, doi: 10.1145/3236391.
- [9] M. Al-Sarem, A.-H. Emar, and A. A. Wahab, "Performance of authorship attribution classifiers with short texts: application of religious Arabic fatwas," *International Journal of Data Mining, Modelling and Management*, vol. 12, no. 3, pp. 350–364, Jul. 2020, doi: 10.1504/IJDDMM.2020.10030684.
- [10] A. Rabab'Ah, M. Al-Ayyoub, Y. Jararweh, and M. Aldwairi, "Authorship attribution of Arabic tweets," in *2016 IEEE/ACS 13th International Conference of Computer Systems and Applications (AICCSA)*, 2016, pp. 1–6, doi: 10.1109/AICCSA.2016.7945818.
- [11] M. Al-Ayyoub, Y. Jararweh, A. Rabab'ah, and M. Aldwairi, "Feature extraction and selection for Arabic tweets authorship authentication," *Journal of Ambient Intelligence and Humanized Computing*, vol. 8, no. 3, pp. 383–393, 2017, doi: 10.1007/s12652-017-0452-1.



- [12] H. N. Alsager, "Towards a stylometric authorship recognition model for the social media texts in Arabic," *Arab World English Journal (AWEJ)*, vol. 11, no. 4, pp. 490–507, Nov. 2021, doi: 10.24093/awej/vol11no4.31.
- [13] A. E. Kah and I. Zeroual, "An empirical evaluation of ensemble bagging-based model for authorship attribution on Twitter," in *2021 Fifth International Conference on Intelligent Computing in Data Sciences (ICDS)*, pp. 1–5, Oct. 2021, doi: 10.1109/ICDS53782.2021.9626735.
- [14] M. Al-Sarem, F. Saeed, A. Alsaedi, W. Boulila, and T. Al-Hadhrami, "Ensemble methods for instance-based arabic language authorship attribution," *IEEE Access*, vol. 8, pp. 17331–17345, Jan. 2020, doi: 10.1109/ACCESS.2020.2964952.
- [15] Y. Addabe, Y. Abu Hammad, N. Ayyad, and A. Yahya, "A dataset for authorship analysis of short modern Arabic text," *Graduation Project, Department of Electrical and Computer Engineering*, Birzeit University, Feb. 2021. (Available from <https://fada.birzeit.edu/handle/20.500.11889/6743>)
- [16] K. Abainia, S. Ouamour, and H. Sayoud, "A novel robust Arabic light stemmer," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 29, no. 3, pp. 557–573, May 2017, doi: 10.1080/0952813X.2016.1212100.
- [17] K. Taghva, R. Elkhoury, and J. Coombs, "Arabic stemming without a root dictionary," in *International Conference on Information Technology: Coding and Computing (ITCC '05)*, vol. 2, pp. 152–157, May 2005, doi: 10.1109/ITCC.2005.90.
- [18] A. Pasha *et al.*, "Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of arabic," in *Proceedings of the ninth international conference on language resources and evaluation (LREC'14)*, May 2014, pp. 1094–1101.
- [19] A. E. Kah and I. Zeroual, "The effects of pre-processing techniques on Arabic text classification," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 10, no. 1, pp. 41–48, 2021, doi: 10.30534/ijatcse/2021/061012021.
- [20] A. E. Kah and I. Zeroual, "Improved document categorization through feature-rich combinations," in *Proceedings of the International Conference on Artificial Intelligence and Computer Vision (AICV2021)*, Springer, Cham, Jun 2021, pp. 346–355, doi: 10.1007/978-3-030-76346-6_32.
- [21] D. Namly, K. Bouzoubaa, A. E. Jihad, and S. L. Aouragh, "Improving Arabic lemmatization through a lemmas database and a machine-learning technique," *Recent Advances in NLP: The Case of Arabic Language*. Studies in Computational Intelligence, Springer, Cham, vol. 874, pp. 81–100, 2020, doi: 10.1007/978-3-030-34614-0_5.
- [22] R. S. Baraka, S. Salem, M. Abu Hussien, N. Nayef, and W. Abu Shaban, "Arabic text author identification using support vector machines," *Journal of Advanced Computer Science and Technology Research*, vol. 4, no. 1, pp. 1–11, Mar. 2014.
- [23] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky, "The stanford CoreNLP natural language processing toolkit," in *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, Jun 2014, pp. 55–60, doi: 10.3115/v1/P14-5010.
- [24] R. Ramezani, "A language-independent authorship attribution approach for author identification of text documents," *Expert Systems with Applications*, vol. 180, p. 115139, Oct. 2021, doi: 10.1016/j.eswa.2021.115139.
- [25] A. Omar and W. I. Hamouda, "The effectiveness of stemming in the stylometric authorship attribution in Arabic," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 1, pp. 116–121, 2020, doi: 10.14569/IJACSA.2020.0110114.

BIOGRAPHIES OF AUTHORS





Dr. Anoual El Kah    holds a Ph.D. degree, since 2019, in Computer Science from Mohamed First University, Morocco with specialization in Computer Science. She primarily works and has authored several articles and book chapters on Language Teaching and Learning, Text Classification, Natural Language Processing, and Machine Learning. She can be contacted at email: elkah.anoual.mri@gmail.com.



Mr. Aymane El Airej   received an MSc degree from Faculty of Sciences and Technics Errachidia in 2021. He is currently a PhD student at Moulay Ismail University of Meknes, Morocco. He is also a member team at ISIC-TEAM ESTM, (L2ISEI)-Laboratory FS in the same university. He primary works on big data and predictive analytics. He can be contacted at email: aymane.airej@gmail.com.



Prof. Dr. Imad Zeroual   is currently an assistant professor with the department of computer science at Faculty of Sciences and Technics, Moulay Ismail University, Morocco. He holds a Ph.D. degree in Computer Science from Mohamed First University with specialization in Artificial Intelligence and Data Science. He primarily works on Natural Language Processing, Machine Learning, Information Retrieval, and Language Teaching and Learning. Prof. Zeroual is a member of various international associations such as the International Association for Educators and Researchers and the International Association of Engineers. He has authored many scientific papers and book chapters with reputed publishers such as Elsevier, Springer, and IEEE. He also served as a reviewer and guest editor of several journals and international conferences. He can be contacted at email: i.zeroual@umi.ac.ma.