Bol. Soc. Paran. Mat.
©SPM – E-ISSN-2175-1188
SPM: www.spm.uem.br/bspm

(3s.) **v. 2025 (43) 7** : 1–10.
ISSN-0037-8712

Towards Robust Arabic Authorship Attribution: A Transformer-Based Model for Multi-Author Imbalanced Corpora

Salah Khennouf, Hassina Hadjadj, Mounir Bouras, Abdelhafid Benyounes and Halim Sayoud

ABSTRACT: Authorship Attribution (AA) seeks to identify the author of a text by analyzing distinctive linguistic and stylistic features. While several studies have focused on English and other Latin-based languages, Arabic AA -particularly with imbalanced datasets- remains comparatively underexplored. In order to close this gap, this work uses a newly created dataset entitled SAB-2, which consists of 7 Arabic books with varying segment lengths, to examine the effectiveness of Arabic pretrained transformer models for multi-class AA. Given the strong impact of data imbalance on classification performance, we apply the Synthetic Minority Oversampling Technique (SMOTE) to enhance minority-class representation and examine its influence on model accuracy. Our experiments evaluate several transformer-based models -AraBERT, AraELECTRA, ARBERT, and MARBERT- alongside deep learning architectures (LSTM, CNN, and a hybrid LSTM-CNN model). Results show that SMOTE substantially improves performance across all models, with the LSTM-CNN architecture combined with AraBERT achieving the highest accuracy of 89%, outperforming baseline experiments without balancing. The obtained results show the robustness of Arabic pretrained transformers in capturing stylistic features from limited and imbalanced textual data, highlighting their potential for advancing Arabic AA in resource-constrained domains.

Key Words: Arabic authorship attribution, Transformer-based models, SMOTE, deep learning for NLP, imbalanced datasets, stylometry.

Contents

1. Introduction

Authorship Attribution (AA) is a long-standing task in computational linguistics concerned with identifying the writer of a given text through the analysis of linguistic and stylistic cues. With the rise of digital communication and the spread of online textual content, AA has become increasingly relevant for applications ranging from digital forensics and cybercrime analysis to plagiarism detection and intellectual property protection.

While substantial progress has been made in authorship analysis for English and other Indo-European languages, Arabic authorship attribution (AAA) remains greatly underexplored due to the linguistic complexity of Arabic, scarcity of labeled resources, and challenges of morphological richness [1].

Recent years have witnessed growing interest in developing transformer-based models tailored to Arabic, which have significantly advanced multiple NLP tasks owing to their ability to capture contextual and semantic dependencies [2]. Arabic variants of BERT and ELECTRA (AraBERT, ARBERT, MARBERT, and AraELECTRA) have revealed strong performance across sentiment analysis, question answering, and sequence labeling. However, their potential for AA in real-world scenarios, especially under imbalanced or limited data conditions, remains insufficiently investigated [3].

Existing AA research has predominantly focused on balanced datasets, with limited attention to scenarios where author contributions vary significantly—a common issue in literary and forensic datasets. This gap is further pronounced for Arabic, where datasets often contain heterogeneous text lengths and uneven author representation. Emerging studies on Arabic poetry and classical literature authorship have begun to demonstrate the value of transformer models for stylistic discrimination yet comprehensive evaluations across multiple Arabic transformer architectures for multi-class AA remain lacking [4].

In light of these challenges, this study explores the effectiveness of several Arabic pretrained transformer models—namely AraBERT, ARBERT, MARBERT, and AraELECTRA—for multi-class AA using an imbalanced Arabic dataset (SAB-2) constructed from seven books of varying segment lengths. To mitigate data imbalance, we apply the Synthetic Minority Oversampling Technique (SMOTE) and examine its impact on classification performance. By comparing transformers with traditional Deep Learning Architectures (DLA), including LSTM, CNN, and a hybrid LSTM-CNN model, this research provides new empirical insights into the robustness of Arabic pretrained language models under real-world constraints.

2. Related Works

The AA has been widely explored within computational linguistics and digital humanities, evolving significantly from traditional stylometric techniques to modern Deep Learning (DL) and transformer-based approaches. Early work in stylometry focused primarily on identifying consistent linguistic patterns that characterize an author's style, regardless of topic or genre. Stamatatos (2009) provided one of the most influential surveys in the field, outlining classical features such as word frequency distributions, character n-grams, syntactic patterns, and lexical richness indicators [5].

These early approaches relied heavily on handcrafted features and statistical machine learning algorithms such as Support Vector Machines, Naïve Bayes, and decision trees, which demonstrated promising performance on long, homogeneous texts. In the context of Arabic, several studies have adapted stylometric methods to the linguistic particularities of the language.

AlZahrani et al.(2023) examined the AA across modern and classical Arabic corpora and showed that morphological and orthographic features—such as affixes, diacritics, and clitic structures—play a critical role in distinguishing authors [6].

Similarly, Alsager (2020) expanded the stylometric feature set to include morphological templates and root-based patterns, improving classification performance on contemporary Arabic writings. However, these feature-engineered approaches are often limited by their sensitivity to topic variation, text length constraints, and the inherent complexity of Arabic morphology [7].

With the rise of DL, researchers shifted toward neural frameworks capable of learning stylistic representations automatically without explicit feature engineering. Early neural models included Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), which were applied to English, Arabic, and multilingual corpora. Shrestha et al. (2017) applied character level CNNs for AA on short texts and demonstrated that CNNs trained on character n grams can effectively capture stylistic cues and perform competitively in attribution tasks, showing robustness to lexical variability compared with traditional feature based models [8].

For Arabic, Mohammed & Kora (2022) demonstrated that DLA such as Bi-LSTMs and CNNs outperform classical methods, particularly when dealing with large-scale and heterogeneous datasets [9]. Almarwani & Diab 2017 further highlighted the advantage of character-level representations over word-level embeddings in Arabic authorship tasks due to the language's rich morphology and complex orthography [10].

More recently, transformer-based models have reshaped the landscape of the AA by leveraging attention mechanisms and contextual embeddings. BERT and its variants have proven to be highly effective in

capturing subtle stylistic patterns that extend beyond surface-level features. Fabien et al 2020. demonstrated that fine-tuned BERT models outperform traditional stylometric baselines, particularly when texts vary in topic or structure [11]. Boenninghoff et al 2021. further argued that transformer representations are capable of preserving authorial style even in cross-topic and adversarial settings [12].

In the Arabic domain, pretrained language models such as AraBERT [13], MARBERT [14] and AraELECTRA [15] have shown exceptional performance in various NLP tasks, motivating researchers to adapt them for authorship classification. While transformer-based AA has advanced considerably in English and other Indo-European languages, it remains relatively underexplored in Arabic, leaving room for contributions that leverage contextual embedding techniques on Arabic corpora.

Another relevant strand of research concerns class imbalance, a common issue in authorship datasets, especially when text availability varies widely between authors. The SMOTE, introduced by Chawla, et al. (2002) [16], has been widely adopted to enhance classification performance by generating synthetic samples for minority classes. More recent works applied SMOTE and its extensions to text classification and AA. Studies such as Kadhim (2019), on SMOTE and its variants, have shown that balancing imbalanced datasets can improve the stability of DL classifiers and mitigate bias toward majority classes —a common challenge when dealing with limited or uneven author contributions [17].

Taken together, the literature reveals a clear evolution in the field of AA: from handcrafted stylometric features toward data-driven DL models and, ultimately, to sophisticated transformer-based architectures. While significant progress has been made, particularly in English and other widely studied languages, AAA remains comparatively understudied, with existing efforts limited in scale, diversity, and methodological depth.

This gap highlights the need for more comprehensive datasets and more robust machine learning frameworks tailored to the linguistic complexity of Arabic. In this context, the present work contributes by exploring transformer-based AA enhanced with SMOTE-based balancing techniques, applied to a newly constructed Arabic corpus designed specifically for evaluating AA performance.

3. Materials and Methods

3.1. Dataset Design and Segment Structure

We have developed a new AAA dataset, which collects seven Arabic books and their authors (7 authors writing on religious topics). This dataset is referred to as SAB-2 (Seven Arabic Books). The books are partitioned into divergent text segments, taking different lengths, ranging within an interval and each one having as an average length of 2900 words. Actually, prior research [18]has demonstrated that in order to guarantee a strong authorship performance, each text should be at least 2500 words in length. These data are written in table 1.

Table 1: SAB-2 dataset description.

Authors	**Segments**	**parameter#**
Hassan's book	29 segments	Big
Alarifi's book	8 segments	Small
AlGhazali's book	39 segments	Big
AlQuaradhawi's book	13 segments	Small
Abdelkafy's book	10 segments	Small
Alkarny's book	23 segments	Big
Amro-khaled's book	9 segments	Small

Big and small are logical parameters (binary value).

3.2. Transformer-Based Authorship Attribution Enhanced with SMOTE

This work aims to improve the AA rate in the presence of a class imbalance. In the following, we introduce an SMOTE oversampling layer to rebalance the class distributions and reduce the majority class bias (see Figure 1). The approach enhances a better understanding of minority classes, which leads to more stable and robust overall model performance.

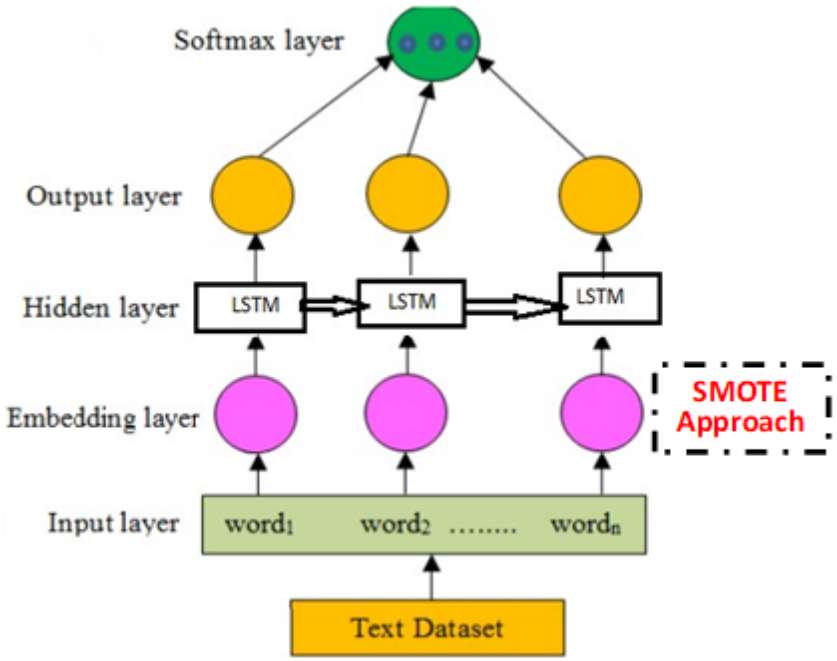

Figure 1: General scheme of the proposed approach

3.2.1. Text preprocessing. The textual data needs to be pre-processed in order to enhance data quality and identification performance. In this sense, numbers, diacritical marks, punctuation and Latin letters, are eliminated from the texts. Next, UTF8 encoding is applied to each text.

3.2.2. Pre-trained Language Models (PLMs). Pre-trained Language Models (PLMs) are Deep Learning Models (DLMs) that are first trained on large collections of text to learn general patterns of language, such as grammar, semantics, and context. After this pretraining, they can be fine-tuned for specific tasks like text classification, sentiment analysis, or authorship attribution. On various downstream tasks, pretraining transformer-based models, such as BERT-based and ELECTRA-based models, demonstrated a notable result [19]. For that reason, we investigated several base pretrained transformers such as: AraELECTRA, AraBERT, ARBERT, and MARBERT.

A replaced token detection goal is used by AraELECTRA [13] on a big Arabic textual dataset. The latter outperforms existing SOTA Arabic language representation models, according to performance results on a number of downstream NLP tasks. When compared to multilingual BERT, AraBERT, an Arabic pretrained model, attained SOTA performance on Arabic NLP tasks [2]. AraBERT was evaluated on a various tasks, comprising question-answering, sentiment analysis, and named entity recognition. AraBERTv0.1-base, AraBERTv0.2base, AraBERTv0.2-large, AraBERTv1-base, AraBERTv2-base, and AraBERTv2-large are the six model versions of the AraBERT. For a variety of Arabic NLP tasks, AraBERT is openly accessible. Models based on BERT, ARBERT and MARBERT, were introduced by Abdul-Mageed, (2020) [14].

In order to train ARBERT, 61 GB (6.5 B tokens) of contemporary standard Arabic text were collected from books, news articles, Wikipedia, and crawled data. A total of 128 GB of Tweets containing at least three Arabic words from different Arabic dialects were used to train MARBERT. With minimal preprocessing, the Tweets were preserved in their original format. The AraBERT model's architecture for AA tasks is shown in Figure 2 below.

3.2.3. Synthetic Minority Oversampling Technique (SMOTE). In recent years, class imbalance has become a critical challenge in data classification, as skewed distributions between majority and minority classes can lead to biased model performance. Sampling methods are commonly used to address this issue

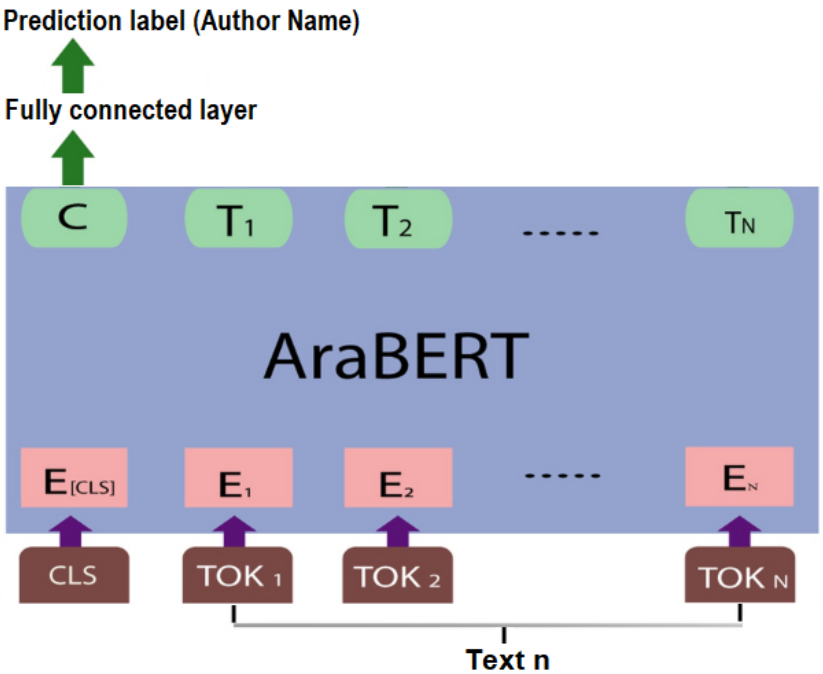

Figure 2: The AraBERT model architecture for AA tasks (adapted from R. Bensoltane & T. Zaki, 2022)

by modifying the class distributions in the training dataset to achieve a more balanced representation [21].

These methods broadly fall into two categories: undersampling, which reduces the number of samples in the majority class, and oversampling, which increases the number of samples in the minority class. Among the oversampling techniques, the SMOTE is widely recognized for its effectiveness. SMOTE, introduced by Chawla et al. (2002) [16], generates synthetic samples of the minority class by interpolating between existing instances, thereby enhancing model learning and mitigating the risk of overfitting.

The SMOTE process generates new virtual samples referring to the similarities in the feature space among existing minority examples. The values of these instances are derived using interpolation and not extrapolation, in order to ensure relevance to the underlying dataset.

Explicitly, SMOTE generates attribute values for new data instances and uses a k-nearest neighbor technique to interpolate values for each minority class instance. For a particular k, the k-nearest neighbors are presented as the k elements of a minority class samples whose Euclidian distance shows the lowest magnitude along the n-dimension of feature space.

To make an artificial instance, we choose one of the k-nearest neighbors at random, multiply the matching feature vector difference by a number between 0 and 1 randomly, and then add this vector to the instance as indicated in Equation 1.

$$y_{\text{new}} = y_i + (y_i' - y_i) \times \delta$$

Where;
y_{new} denotes a new instance,
y_i is the minority under consideration,
y_i' is one of the k-nearest neighbors for y_i,
δ is a random value between 0 and 1.

In ML and Data Mining, SMOTE is regarded as an important data sampling or preprocessing approach [22]. We employed SMOTE in the present work because of its widespread use and impact. As illustrated in Figure 3, oversampling and undersampling are commonly employed techniques to address class imbalance.

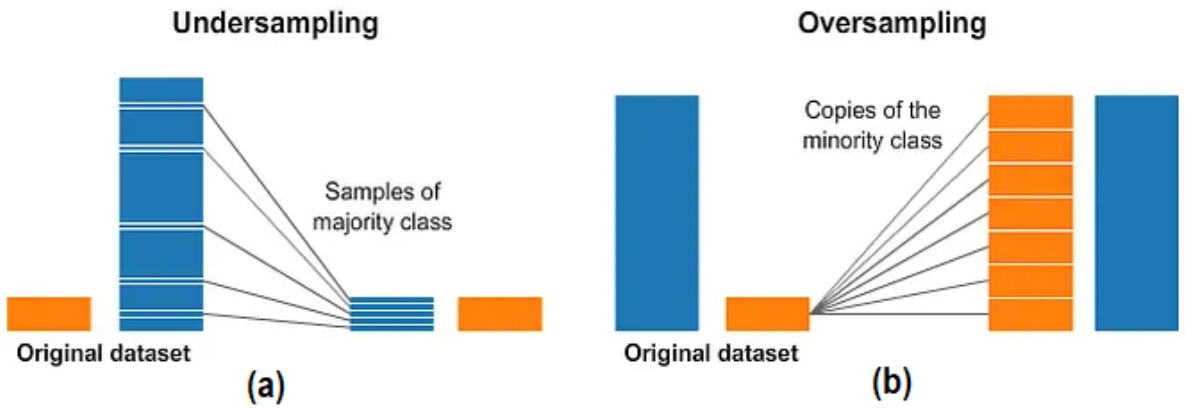

Figure 3: Handling class imbalance: (a) Undersampling, (b) Oversampling. (Adapted from Brownlee, 2020).

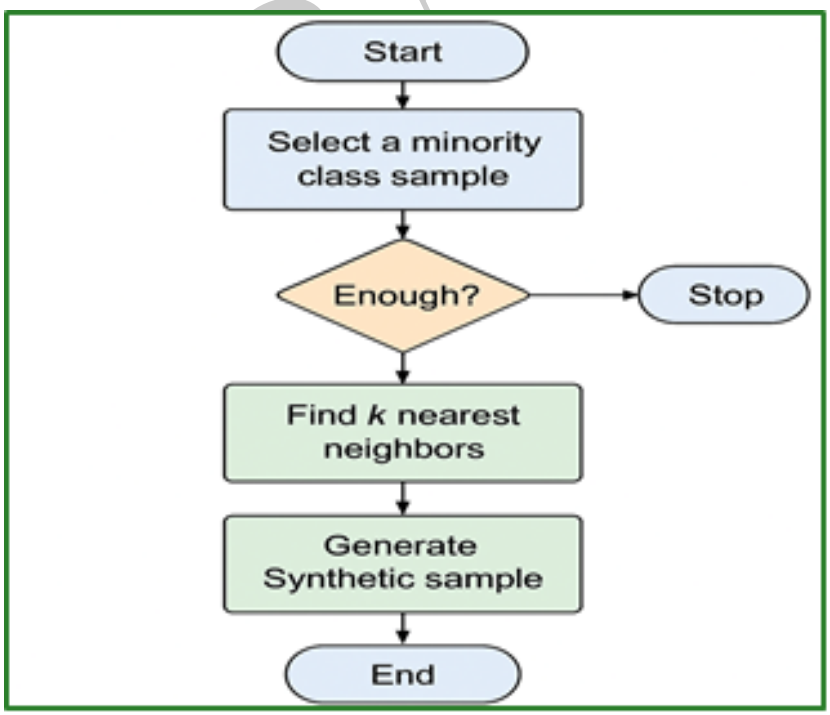

Figure 4: Flowchart of the SMOTE Approach (adapted Chawla et al., 2002)

3.3. Deep Learning Models (DLMs)

The DL techniques are widely used in the field of text classification, particularly for authorship attribution; among these techniques are Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). In our experiments, the authorship is classified using three models (LSTM, CNN and LSTM-CNN). The description of the models is as follows:

3.3.1. Long Short-Term Memory (LSTM). Logically, RNN can be used to justify Long Short-Term Memory Network (LSTM) [23]. Consequently, LSTM is a category of RNN that can learn and retain patterns across lengthy sequences. For tasks requiring sequential data, like Time Series Analysis (TSA) and Natural Language Processing (NLP), this makes it highly effective.

In contrast to conventional RNNs, LSTM is able to learn from data with significant gaps between pertinent events because of a special design that prevents the "vanishing gradient" issue. In this work, an embedding layer, an LSTM layer, a dropout layer, an additional LSTM layer, and a dense layer were used to built our LSTM model.

3.3.2. Convolutional Neural Networks (CNNs). The most common use of CNNs, a well-known class of DLMs, is for visual data analysis [24]. For example, they are prepared to automatically and adaptively learn spatial hierarchies of features from tasks like video and image classification.

One or more convolutional layers make up CNNs, which are frequently followed by pooling, fully connected, and normalization layers. An embedding layer, three sets of Conv1D and MaxPooling1D layers, a dropout layer following each MaxPooling1D layer, a GlobalMaxPooling1D layer, and a dense layer were used to build our CNN model.

3.4. Evalution metrics

Evaluation metrics are crucial for evaluating the performance of the classification. Accuracy is one of the most widely used measures, even in imbalanced datasets. In the present work, the accuracy is calculated using the following formula:

$$\text{Accuracy} = \frac{\text{Number of correctly classified segments}}{\text{Total number of tested examples}} \tag{3.1}$$

It would be challenging to provide enough data to separate between the training and testing sets in a classification problem with limited dataset. Thus, the n-fold cross validation technique can be applied in this case [25]. This procedure is generally applied in ML models, to significantly evaluate the algorithms using the overall dataset (training and testing). Five-fold cross validation were used in this work.

In summary, the data was split into 5 equal-sized sets at random. Four partitions for training the authorship model, and the fifth partition is used for test. After that, the process is recurrent with every fold being detained for examination. Thus, classification is carried out five times, using four partitions for training and one different partition for testing each time. Finally, to determine the mean results for the dataset, the outcomes of the classification tasks are then combined.

4. Results and Discussion

In this inquiry, we conducted several experiments on Arabic AA using Transformer-based Models for multi-author on unbalanced corpora. The main purpose was to assess the robustness and effectiveness of pretrained transformer-based models for author identification in specialized domains with small and unbalanced dataset.

First, to deal with this issue of unbalanced data, the SMOTE technique was applied to generate a new dataset with an equal number of samples across all categories. Figure 5 illustrates examples of artificially generated data distributions across seven authors after balancing the dataset.

The conducted experiments showed the effectiveness of the usage of SMOTE and Transformer-based Models in AAA on imbalanced datasets. Table 2 & Figure 6 illustrate the results of the comparison between different models with and without balancing techniques.

As shown, the application of SMOTE improves the accuracy values obtained for all models and transformer architectures. The LSTM-CNN model combined with AraBERT achieves the highest accuracy of

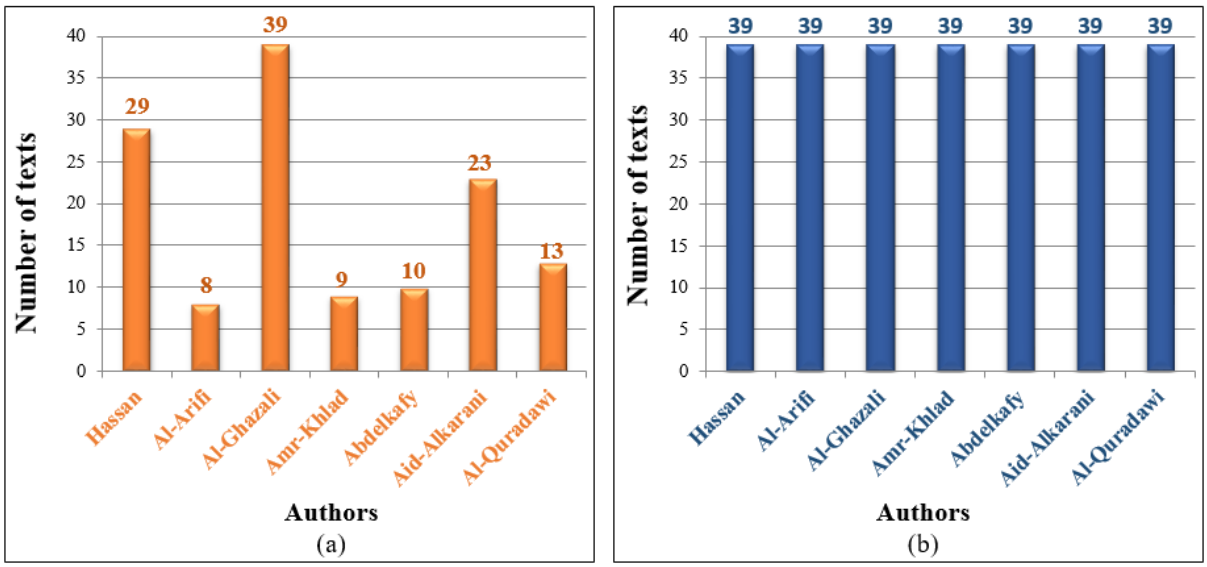

Figure 5: Text samples distribution among authors. (a) : Original text samples. (b) : Balanced text samples using SMOTE approach

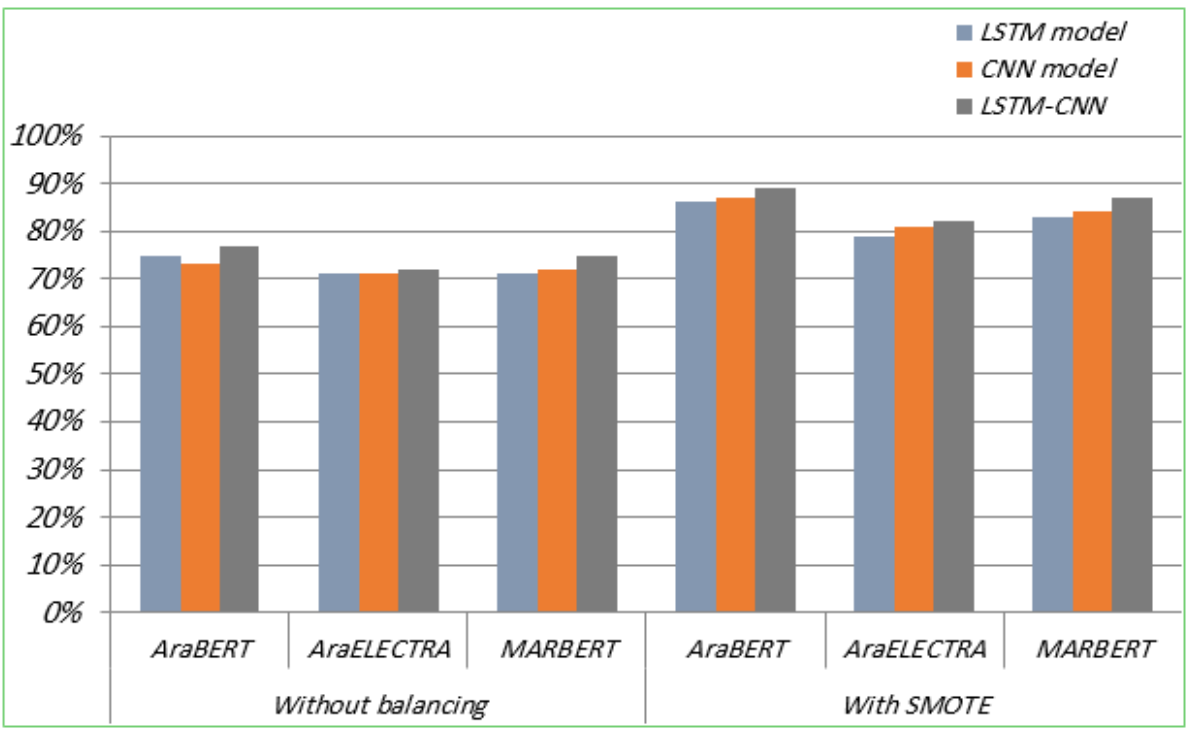

Figure 6: Comparison between our methods. Left: without balancing. Right: with SMOTE

Table 2: Accuracies obtained without balancing and with balancing using SMOTE approach

Accuracy	Accuracy Without balancing			Accuracy With SMOTE		
	AraBERT	AraELECTRA	MARBERT	AraBERT	AraELECTRA	MARBERT
LSTM model	75%	71%	71%	86%	79%	83%
CNN model	73%	71%	72%	87%	81%	84%
LSTM-CNN model	77%	72%	75%	89%	82%	87%

89% using SMOTE approach, compared to 77% without balancing. This represents a 12% improvement in performance. Similarly, all other models show considerable improvements ranging from 8% to 14% when SMOTE is applied.

These results, indicate that Arabic pretrained transformer models, particularly AraBERT, are highly effective for authorship attribution tasks even with imbalanced data. The combination of deep learning architectures with transformer embeddings and SMOTE-based balancing creates a robust framework for AAA that can handle real-world data distribution challenges.

Furthermore, the LSTM-CNN model combined with AraBERT demonstrates superior performance compared to the other models using different pretrained Transformers.

5. Conclusion

The relevance of the pretrained transformer-based models for AAA tasks in specialized domains, such as theological law, which are frequently limited by sparse and unbalanced data, is highlighted in this study. By fine-tuning multiple models on texts from authors sharing similar historical and linguistic contexts, we demonstrated that these models can effectively capture stylistic and semantic nuances beyond traditional feature-based approaches.

The evaluated models (AraBERT) and balancing approach (SMOTE) consistently achieved superior performance, suggesting its robustness for Arabic AA tasks. These findings open the door for further exploration of transformer-based architectures in specialized domains and under resource-constrained conditions.

As future work, we plan to extend our experiments to larger and more highly imbalanced corpora in order to obtain better results.

References

1. Jambi, K. M., Khan, I. H., Siddiqui, M. A., & Alhaj, S. O. (2021). Towards authorship attribution in arabic short-microblog text. IEEE Access, 9, 128506-128520.

2. Antoun, W., Baly, F., & Hajj, H. (2020a). Arabert: Transformer-based model for arabic language understanding. ArXiv Preprint ArXiv:2003.00104.

3. Mashaabi, M., Al-Khalifa, S., & Al-Khalifa, H. (2024). A Survey of Large Language Models for Arabic Language and its Dialects. arXiv preprint arXiv:2410.20238.

4. Alqurashi, L., Sharoff, S., Watson, J., & Blakesley, J. (2025, January). BERT-based Classical Arabic Poetry Authorship Attribution. In Proceedings of the 31st International Conference on Computational Linguistics (pp. 6105-6119).

5. Stamatatos, E. (2009). A survey of modern authorship attribution methods. Journal of the American Society for information Science and Technology, 60(3), 538-556.

6. AlZahrani, F. M., & Al-Yahya, M. (2023). A transformer-based approach to authorship attribution in classical arabic texts. Applied Sciences, 13(12), 7255.

7. Alsager, H. N. (2020). Towards a Stylometric Authorship Recognition Model for the Social Media Texts in Arabic. Arab World English Journal, 11(4), 490-507.

8. Shrestha, P. et al. (2017). Convolutional neural networks for authorship attribution of short texts. In Proceedings of the 15th conference of the European chapter of the association for computational linguistics: Volume 2, short papers (pp. 669-674).

9. Mohammed, A., & Kora, R. (2022). An effective ensemble deep learning framework for text classification. Journal of King Saud University-Computer and Information Sciences, 34(10), 8825-8837.

10. Almarwani, N., & Diab, M. (2017, April). Arabic textual entailment with word embeddings. In Proceedings of the third arabic natural language processing workshop (pp. 185-190).

11. Fabien, M., Villatoro-Tello, E., Motlicek, P., & Parida, S. (2020, December). BertAA: BERT fine-tuning for Authorship Attribution. In Proceedings of the 17th International Conference on Natural Language Processing (ICON) (pp. 127-137).

12. Boenninghoff, B., Nickel, R. M., & Kolossa, D. (2021). O2D2: Out-of-distribution detector to capture undecidable trials in authorship verification. arXiv preprint arXiv:2106.15825.

13. Antoun, W., Baly, F., & Hajj, H. (2020b). AraELECTRA: Pre-training text discriminators for Arabic language understanding. ArXiv Preprint ArXiv:2012.15516.

14. Abdul-Mageed, M., & Elmadany, A. (2021, August). ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers) (pp. 7088-7105).

15. Antoun, W., Baly, F., & Hajj, H. (2021, April). AraELECTRA: Pre-training text discriminators for Arabic language understanding. In Proceedings of the sixth arabic natural language processing workshop (pp. 191-195).

16. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research. https://doi.org/10.1613/jair.953

17. Kadhim, A. I. (2019). Survey on supervised machine learning techniques for automatic text classification. Artificial intelligence review, 52(1), 273-292.

18. Eder, M. (2015). Does size matter? Authorship attribution, small samples, big problem. Digital Scholarship in the Humanities. https://doi.org/10.1093/llc/fqt066

19. Alrowili, S., & Vijay-Shanker, K. (2021, November). ArabicTransformer: Efficient large Arabic language model with funnel transformer and ELECTRA objective. In Findings of the association for computational linguistics: EMNLP 2021 (pp. 1255-1261).

20. Abdul-Mageed, M., Elmadany, A., & Nagoudi, E. M. B. (2020). ARBERT & MARBERT: Deep bidirectional transformers for Arabic. ArXiv Preprint ArXiv:2101.01785.

21. Hoang, G., Bouzerdoum, A., & Lam, S. (2009). Learning Pattern Classification Tasks with Imbalanced Data Sets. In Pattern Recognition. https://doi.org/10.5772/7544

22. García, S., Luengo, J., & Herrera, F. (2016). Tutorial on practical tips of the most influential data preprocessing algorithms in data mining. Knowledge-Based Systems. https://doi.org/10.1016/j.knosys.2015.12.006

23. Gungor, A. (2018). Fifty Victorian Era Novelists Authorship Attribution Data.

24. Machicao, J., Corrêa Jr, E. A., Miranda, G. H. B., Amancio, D. R., & Bruno, O. M. (2018). Authorship attribution based on life-like network automata. PloS One, 13(3), e0193703.

25. Allgaier, J., & Pryss, R. (2024). Cross-validation visualized: a narrative guide to advanced methods. Machine Learning and Knowledge Extraction, 6(2), 1378-1388.

Salah Khennouf, University of M'sila, M'sila, Algeria. Orcid number of the first author
E-mail address: salah.khennouf@univ-msila.dz

and

Hassina Hadjadj, University of Science and Technology – Houari Boumediene, Bab Ezzouar, Algeria. Orcid number of the second author
E-mail address: hadjadj.has@gmail.com

and

Mounir Bouras, University of M'sila, M'sila, Algeria. Orcid number of the third author
E-mail address: mounir.bouras@univ-msila.dz

and

Abdelhafid Benyounes, University of M'sila, M'sila, Algeria. Orcid number of the fourth author
E-mail address: abdelhafid.benyounes@univ-msila.dz

and

Halim Sayoud, University of Science and Technology – Houari Boumediene, Bab Ezzouar, Algeria. Orcid number of the fifth author
E-mail address: halim.sayoud@gmail.com