Performance Analysis of Multi-Model Transformer Architectures for classical Arabic Author Identification

Mayar Boghdady , Ahmed Saeed , Yomna Eid ,Ensaf Mohamed
CIS , School of Information Technology and Computer Science ,Nile University
{MBoghdadi,ahsaeed,YEid,EnMohamed*}*@nu.edu.eg

Abstract

Authorship identification for Arabic-script languages presents unique challenges due to rich morphology, ortho graphic variation, and limited annotated resources. This pa per presents a transformer-based system developed for the AbjadAuthorID [1] shared task, which addresses multiclass authorship identification for Arabic-script texts. We fine tune AraBERTv2 [2] using a standardized preprocessing pipeline designed to normalize Arabic orthography. The ex periments were conducted on a large-scale dataset compris ing 21 authors and more than 35K text segments. The pro posed approach achieves a weighted F1-score of 91% on the validation set and 87% on the official test set, demonstrating the effectiveness of transformer-based language models for authorship attribution in Arabic-script languages.

1. Introduction

Languages written in Arabic-derived scripts form one of the most linguistically diverse and culturally rich language groups worldwide. Although these languages are used by nearly one billion speakers, many remain under-resourced in Natural Language Processing (NLP), particularly in tasks requiring annotated datasets and robust linguistic tools. Au thorship identification, the task of determining the author of a given text, is a fundamental NLP problem with applica tions in plagiarism detection, literary analysis, digital foren sics, and digital humanities.

The AbjadNLP workshop [1] aims to advance research and resources for Arabic and other Arabic-script languages through inclusive and community-driven efforts. Within this initiative, the AbjadAuthorID shared task focuses on multi class authorship identification, introducing additional chal lenges such as stylistic similarity across authors, histori cal and literary variations, and orthographic inconsistencies. These characteristics make robust

modeling and effective preprocessing essential.

In this paper, we present a transformer-based solution leveraging AraBERT v2, adapted to Arabic linguistic prop erties via task-specific fine-tuning and orthographic normal ization. The main contributions of this work are summarized as follows:

- A reproducible AraBERT v2 fine-tuning pipeline for multiclass authorship identification in Arabic.

- A detailed preprocessing workflow tailored for Arabic orthography and noise reduction.

- Experimental results on the AbjadAuthorID dataset demonstrating competitive performance.

This work is submitted to the AbjadAuthorID [1] shared task at the 2nd Workshop on NLP for Languages Using Arabic Script. By addressing the specific complexities of classical Arabic and authorial style, we aim to contribute to the workshop's goal of advancing robust, inclusive NLP re sources for Arabic-script languages.

2. Related Work

Authorship identification (also referred to as authorship attribution) is a core task in stylometry and computational linguistics that aims to identify the author of a document based on writing style. Early research in this area relied pri marily on feature engineering approaches, extracting lexical, syntactic, and structural cues such as character and word *n* grams, function-word frequencies, and punctuation patterns. These representations were commonly combined with tradi tional machine learning classifiers such as Support Vector Machines (SVM) and Logistic Regression, achieving strong results across various benchmarks, particularly in closed-set settings.

Shared evaluation campaigns have played a significant role in shaping the field of authorship analysis by provid ing standardized tasks and benchmark datasets. Among the most influential initiatives is PAN, which has hosted mul tiple shared tasks that cover authorship attribution, author ship verification, and writing style analysis. PAN bench marks have enabled consistent evaluation and

motivated the development of more robust and transferable approaches in authorship analysis [6, 8, 9].

With the development of deep learning, authorship attri bution research gradually shifted toward neural approaches that reduce the reliance on handcrafted stylometric features. Neural architectures including Convolutional Neural Net works (CNNs) [7] and Recurrent Neural Networks (RNNs), such as Long Short-Term Memory (LSTM) models, have been used to learn latent stylistic representations directly from text. Although neural models can capture complex fea ture interactions, they often require large training corpora and may struggle with generalization when datasets are lim ited or stylistically diverse.

More recently, transformer-based language models have become the dominant approach for text classification and authorship analysis due to their ability to model contex tual dependencies and semantic variation. The BERT ar chitecture [4] has been widely adopted and studied for su pervised fine-tuning. In particular, transformer fine-tuning strategies have been extensively analyzed for text classifica tion tasks, demonstrating that careful selection of hyperpa rameters and training practices can significantly impact per formance [10]. In the context of authorship attribution, fine tuned BERT models have shown strong results, as demon strated by BertAA, which adapts BERT for supervised au thorship classification by adding a task-specific classifica tion layer [5].

Arabic-script authorship identification presents addi tional challenges compared to high-resource Latin-script settings. Arabic is morphologically rich, exhibits flexible word order, and contains substantial orthographic variabil ity. Furthermore, Arabic texts may include dialectal or his torical variation, and may differ in diacritics usage, elonga tion characters, and normalization conventions, all of which affect tokenization and downstream modeling. These lin guistic characteristics motivate the use of robust preprocess ing and language-specific models for Arabic NLP.

To support Arabic NLP, several pretrained transformer models have been proposed, including AraBERT [2], which is pretrained on large-scale Arabic corpora and has consis tently demonstrated strong performance across Arabic lan guage understanding benchmarks. AraBERT is accompa nied by preprocessing utilities tailored for Arabic orthogra phy normalization, making it a strong foundation for down stream classification tasks, including authorship attribution.

Finally, community-driven workshops and shared tasks play an essential role in expanding NLP research beyond high-resource languages. AbjadNLP focuses on Arabic and Arabic-script languages and promotes inclusive, open, and cross-community research efforts. The AbjadAuthorID shared task specifically aims to advance authorship identifi cation in morphologically rich and under-resourced Arabic script settings, encouraging reproducible and scalable ap proaches based on modern pretrained language models [1].

3. Task Description and Dataset

The AbjadAuthorID shared task [1] focuses on multiclass authorship identification for Arabic-script texts. Given a text segment, the objective is to predict its corresponding au thor from a closed set of candidate authors. This setting is challenging due to stylistic similarities between authors, variation in genre and content, and the linguistic complexity of Arabic, including orthographic inconsistencies and rich morphology.
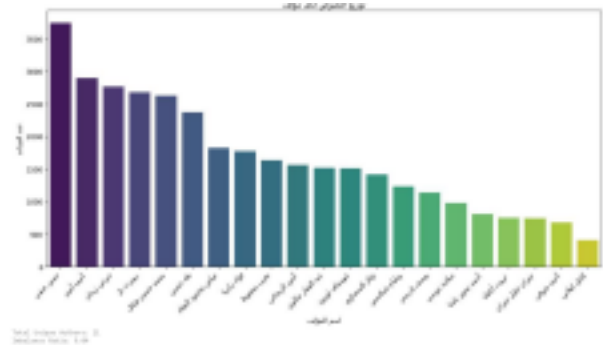

Figure 1. Distribution of training samples across the 21 author classes, showing an imbalanced dataset.

3.1. Dataset

We conduct experiments using the official AbjadAu thorID dataset, which consists of short text segments labeled with one of 21 authors. The dataset is split into training, val idation, and test partitions. The training set contains 35,122 segments, while the validation set contains 4,157 segments. Additionally, we evaluate our final model on the official test set provided by the shared task organizers, which includes 8,413 text segments.

Each instance contains an identifier, the text segment, and the author label. Since the task is formulated as supervised multiclass classification, we map the author names into in teger labels ranging from 0 to 20 to enable model training.

3.2. Class Imbalance

A key characteristic of the dataset is that it is not uni formly distributed across authors. Some authors have sub stantially more segments than others,

resulting in an im balanced class distribution. This imbalance can introduce bias toward majority classes and reduce performance on under-represented authors, especially in a multiclass setting. Therefore, model evaluation is reported using weighted F1-score in addition to accuracy, as it better reflects perfor mance under skewed label distributions.

4. Methodology

We propose a transformer-based authorship identification system built on AraBERTv2 [2], a pretrained Arabic lan guage model. Our approach fine-tunes AraBERTv2 using supervised learning for multiclass classification, leveraging preprocessing techniques designed specifically for Arabic orthography.

4.1. Preprocessing

Arabic texts exhibit significant orthographic variability, including optional diacritics, elongation characters, and in consistent normalization. To reduce noise and improve model robustness, we apply the AraBERT preprocessing pipeline using the AraBERT preprocessor. Specifically, we normalize text by removing diacritics (*tashkeel*) and elonga-

Figure 2. Fine-tuning pipeline using AraBERT v2 for multiclass authorship identification.

tion (*tatweel*), in addition to applying Arabic-specific nor malization rules.

Formally, for each input segment x, we construct a nor malized form $\tilde{x}$:

$$\tilde{x} = \text{Preprocess}(x) \quad (1)$$

where Preprocess($\cdot$) is implemented using ArabertPreprocessor.

4.2. Model Architecture

We fine-tune aubmindlab/bert-base-arabertv2 as the backbone encoder. The representation of the [CLS] token from the final hidden layer is passed to a classification layer predicting one of the 21 author classes. Given an input sequence $\tilde{x}$, the model computes:

$$h = \text{AraBERT}(\tilde{x}) \quad (2)$$

$$\hat{y} = \text{softmax}(Wh_{[CLS]} + b) \quad (3)$$

where $\hat{y}$ is the predicted distribution over authors, and W, b are trainable classification parameters.

4.3. Tokenization

We tokenize each segment using the AraBERT tokenizer with a maximum sequence length of 512 tokens. Segments longer than the limit are truncated. This configuration pro vides a balance between computational efficiency and the ability to capture sufficient context for author-specific pat terns.

5. Experimental Setup

Fine-tuning is conducted using the HuggingFace Trans formers library. Models are trained for three epochs with a batch size of 8 and a learning rate of 2×10^{-5} using AdamW. The best model checkpoint is selected based on the weighted F1-score on the validation set, which is particularly suitable under class imbalance conditions.

5.1. Evaluation Metrics

We report Accuracy and weighted F1-score as primary metrics for multiclass authorship identification. Since the dataset is imbalanced across authors, we additionally re port Macro-F1, which assigns equal importance to each class and provides better insight into performance on under represented authors.

Table 1. Training hyperparameters used in our experiments.

Parameter	Value
Backbone model	AraBERT v2 (base)
Max sequence length	512
Batch size	8
Epochs	3
Learning rate	2×10^{-5}
Optimizer	AdamW
Model selection metric	Weighted F1

Table 2. Performance comparison of AraBERTv2, MARBERT and deepseekv3.2.

Model	Val. Acc.	Val. W-F1	Test W-F1
AraBERT v2	91.07	91.07	87
MARBERT	89	85	84
DeepSeekv3.2	12.53	5.57	

5.2. Training Configuration

We summarize the key training hyperparameters in Ta ble 1.

5.3. Alternative Model and Class Imbalance

In addition to AraBERTv2, this study evaluated two al ternative modeling approaches. First, we employed MAR BERT, a masked language model pretrained on both Mod ern Standard Arabic and dialectal variants. Although MAR BERT demonstrated competitive performance, it was con sistently surpassed by AraBERTv2 across our experimen tal trials. This discrepancy may be attributed to the imbal anced distribution of author classes within the dataset; such data sparsity can introduce bias toward high-frequency au thors while degrading the model's predictive accuracy for minority classes. We also evaluated Zero-shot prompt clas sification using the DeepSeekv3.2 [3] LLM. This approach, relying solely on prompt engineering without fine-tuning, yielded a significantly lower accuracy of 12.53. These re sults confirm that identifying an author's stylistic "finger print" requires task-specific training, as idiosyncratic nu ances—such as syntactic preferences and lexical frequen cies—are too subtle for general-purpose models to extract in zero-shot settings. Consequently, AraBERTv2 [2] remains the more robust paradigm for authorship attribution, partic ularly when the corpus is predominantly composed of Mod ern Standard Arabic.

6. Results

Table 2 presents a comparison between AraBERT v2 and MARBERT. AraBERT achieves superior weighted F1- scores on both validation and test sets, indicating that Arabic-specific preprocessing and AraBERT fine-tuning provide strong modeling of authorial style.

Overall, the results highlight that transformer-based Ara bic language models can effectively capture stylistic features for multiclass authorship identification. The observed per formance gap between validation and test sets suggests that the test partition may contain more challenging samples, in creased stylistic overlap, or higher variability across authors.

7. Error Analysis

Class imbalance remains a significant factor influenc ing system performance; under-represented authors provide fewer training examples, which hinders the model's ability to learn distinctive stylistic patterns. Consequently, predic tions may exhibit a bias toward majority classes, leading to increased confusion between minority authors and stylisti cally similar frequent authors.

Furthermore, text length plays a critical role in attribution accuracy. Short or highly generic segments often lack suf ficient idiosyncratic markers to reliably distinguish an au thor's style. In contrast, longer segments provide the lexi cal and structural density necessary for the model to extract richer cues, ultimately leading to higher confidence in the classification results

8. Conclusion and Future Work

This paper presented a transformer-based system for mul ticlass authorship identification developed for the AbjadAu thorID shared task, held in conjunction with EACL 2026. Our approach fine-tunes AraBERTv2 using Arabic-specific preprocessing for orthographic normalization. Experiments on the AbjadAuthorID dataset demonstrate that pretrained Arabic language models are highly effective for authorship attribution, achieving a weighted F1-score of 91% on the validation set and 87% on the official test set.

In future work, our work will be in different direc tions include longer-context modeling, segment aggrega tion, and interpretability techniques to better understand author-specific stylistic cues learned by transformer models. Finally, extending similar methodologies to other Arabic script languages aligns with the broader goals of AbjadNLP and inclusive language technology.

References

[1] Shadi Abudalfa, Saad Ezzini, Ahmed Abdelali, Mustafa Jar rar, Mo El-Haj, Nadir Durrani, Hassan Sajjad, Farah Adeeba, and Sina Ahmadi. Abjadauthorid: Authorship identification for arabic-script languages at abjadnlp 2026. In *Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script (AbjadNLP 2026), co-located with the 19th Confer ence of the European Chapter of the Association for Com putational Linguistics (EACL 2026)*, Rabat, Morocco, March 2026. 1, 2, 4

[2] Wissam Antoun, Fady Baly, and Hazem Hajj. AraBERT: Transformer-based model for Arabic language understand ing. In Hend Al-Khalifa, Walid Magdy, Kareem Darwish, Tamer Elsayed, and Hamdy Mubarak, editors, *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Pro cessing Tools, with a Shared Task on Offensive Language De tection*, pages 9–15, Marseille, France, may 2020. European Language Resource Association. 1, 2, 3

[3] DeepSeek-AI. Deepseek-v3.2: Pushing the frontier of

open large language models, 2025. 3

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional trans formers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technolo gies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Compu tational Linguistics. 2

[5] Mael Fabien, Esa ̈ u Villatoro-Tello, Petr Motlicek, and Shan- ́ tipriya Parida. Bertaa: Bert fine-tuning for authorship attri bution. In *Proceedings of ICON 2020*, 2020. 2

[6] PAN @ Webis. Pan shared tasks: Authorship analysis and text forensics. https://pan.webis.de/shared tasks.html, 2026. 1

[7] Prasha Shrestha, Sebastian Sierra, Fabio Gonzalez, Manuel ́ Montes, Paolo Rosso, and Thamar Solorio. Convolutional neural networks for authorship attribution of short texts. In Mirella Lapata, Phil Blunsom, and Alexander Koller, editors, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 669–674, Valencia, Spain, Apr. 2017. Association for Computational Linguistics. 1

[8] Efstathios Stamatatos et al. Overview of the authorship ver ification task at pan 2022. In *CLEF Working Notes*, 2022. 1

[9] Efstathios Stamatatos et al. Overview of the authorship ver ification task at pan 2023. In *CLEF Working Notes*, 2023. 1

[10] Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. How to fine-tune bert for text classification? *arXiv preprint arXiv:1905.05583*, 2019. 2