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

Ahmed Saeed, Mayar, Enjy
Nile University

Abstract

Authorship identification for Arabic-script languages presents unique challenges due to rich morphology, orthographic variation, and limited annotated resources. This paper presents a transformer-based system developed for the AbjadAuthorID shared task, which addresses multiclass authorship identification for Arabic-script texts. We fine-tune AraBERT v2 using a standardized preprocessing pipeline designed to normalize Arabic orthography. Experiments are conducted on a large-scale dataset comprising 21 authors and more than 35K text segments. The proposed approach achieves a weighted F1-score of 94% 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. Authorship identification, the task of determining the author of a given text, is a fundamental NLP problem with applications in plagiarism detection, literary analysis, digital forensics, and digital humanities.

The AbjadNLP workshop 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 multiclass authorship identification, introducing additional challenges such as stylistic similarity across authors, historical 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 properties via task-specific fine-tuning and orthographic normalization. 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.

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 primarily 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 traditional 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 providing standardized tasks and benchmark datasets. Among the most influential initiatives is PAN, which has hosted multiple shared tasks that cover authorship attribution, authorship verification, and writing style analysis. PAN benchmarks have enabled consistent evaluation and motivated the development of more robust and transferable approaches in authorship analysis [5–7].

With the development of deep learning, authorship attribution research gradually shifted toward neural approaches that reduce the reliance on handcrafted stylometric features. Neural architectures including Convolutional Neural Networks (CNNs) 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

feature interactions, they often require large training corpora and may struggle with generalization when datasets are limited 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 contextual dependencies and semantic variation. The BERT architecture [3] has been widely adopted and studied for supervised fine-tuning. In particular, transformer fine-tuning strategies have been extensively analyzed for text classification tasks, demonstrating that careful selection of hyperparameters and training practices can significantly impact performance [8]. In the context of authorship attribution, fine-tuned BERT models have shown strong results, as demonstrated by BertAA, which adapts BERT for supervised authorship classification by adding a task-specific classification layer [4].

Arabic-script authorship identification presents additional challenges compared to high-resource Latin-script settings. Arabic is morphologically rich, exhibits flexible word order, and contains substantial orthographic variability. Furthermore, Arabic texts may include dialectal or historical variation, and may differ in diacritics usage, elongation characters, and normalization conventions, all of which affect tokenization and downstream modeling. These linguistic characteristics motivate the use of robust preprocessing 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 consistently demonstrated strong performance across Arabic language understanding benchmarks. AraBERT is accompanied by preprocessing utilities tailored for Arabic orthography normalization, making it a strong foundation for downstream 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 identification in morphologically rich and under-resourced Arabic-script settings, encouraging reproducible and scalable approaches based on modern pretrained language models [1].

3. Task Description and Dataset

The AbjadAuthorID shared task focuses on multiclass authorship identification for Arabic-script texts. Given a text segment, the objective is to predict its corresponding author 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.

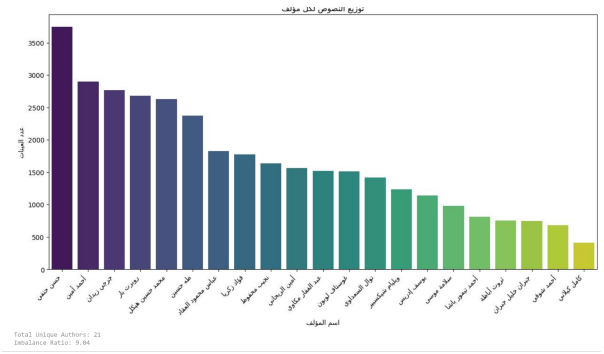

Figure 1. Distribution of training samples across the 21 author classes, showing an imbalanced dataset.

3.1. Dataset

We conduct experiments using the official AbjadAuthorID dataset, which consists of short text segments labeled with one of 21 authors. The dataset is split into training, validation, 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 integer 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 uniformly distributed across authors. Some authors have substantially more segments than others, resulting in an imbalanced 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 performance under skewed label distributions.

4. Methodology

We propose a transformer-based authorship identification system built on AraBERT v2, a pretrained Arabic language model. Our approach fine-tunes AraBERT using supervised learning for multiclass classification, leveraging preprocessing techniques designed specifically for Arabic orthography.

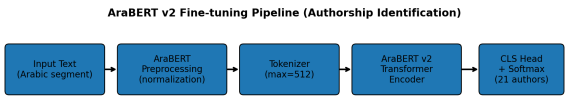

Figure 2. Fine-tuning pipeline using AraBERT v2 for multiclass authorship identification.

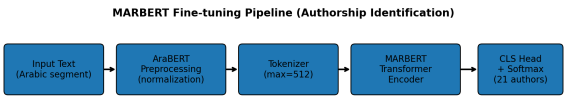

Figure 3. Fine-tuning pipeline using MARBERT for multiclass authorship identification.

4.1. Preprocessing

Arabic texts exhibit significant orthographic variability, including optional diacritics, elongation characters, and inconsistent 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 elongation (*tatweel*), in addition to applying Arabic-specific normalization rules.

Formally, for each input segment x, we construct a normalized form $\tilde{x}$:

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

where $\text{Preprocess}(\cdot)$ 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:

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

$$\hat{y} = \text{softmax}(\mathbf{W}\mathbf{h}_{[\text{CLS}]} + \mathbf{b}) \tag{3}$$

where $\hat{y}$ is the predicted distribution over authors, and $\mathbf{W}, \mathbf{b}$ 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 provides a balance between computational efficiency and the ability to capture sufficient context for author-specific patterns.

5. Experimental Setup

Fine-tuning is conducted using the HuggingFace Transformers 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.

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

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 report Macro-F1, which assigns equal importance to each class and provides better insight into performance on underrepresented authors.

For completeness, we also compute BLEU as an auxiliary metric by comparing the predicted author label (as a text token) with the ground-truth author label. We note that BLEU is not a standard metric for classification tasks and is included only as an additional reference.

5.2. Training Configuration

We summarize the key training hyperparameters in Table 1.

5.3. Alternative Model and Class Imbalance

In addition to AraBERT v2, we experimented with MARBERT, a pretrained masked language model designed to support both Modern Standard Arabic (MSA) and dialectal Arabic. While MARBERT achieved competitive results, it was consistently outperformed by AraBERT v2 in our experiments. One possible reason is the imbalanced distribution of author classes in the dataset, which can bias training toward frequent authors and reduce performance for minority classes. Moreover, AraBERT v2 appears better suited for authorship attribution when the majority of samples are written in MSA.

Table 2. Performance comparison of AraBERT v2 and MARBERT.

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	

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 Arabic language models can effectively capture stylistic features for multiclass authorship identification. The observed performance gap between validation and test sets suggests that the test partition may contain more challenging samples, increased stylistic overlap, or higher variability across authors.

7. Error Analysis

To better understand system limitations, we analyze common sources of misclassification. A major difficulty arises when multiple authors share similar writing style, topic, or genre, reducing the discriminative signal available for classification. This is especially common in literary and philosophical texts where vocabulary and syntactic patterns overlap.

Another important factor is class imbalance. Underrepresented authors contribute fewer training examples, which can make their stylistic patterns harder to learn. As a result, predictions may be biased toward majority classes, increasing confusion between minority authors and stylistically similar frequent authors.

Text length also plays a role. Short or highly generic segments may not contain sufficient author-specific markers to reliably attribute authorship. In contrast, longer segments often provide richer lexical and structural cues and lead to higher confidence predictions.

Future improvements could include imbalance-aware learning strategies such as class-weighted loss functions, oversampling minority authors, or focal loss. Moreover, segment aggregation and longer-context modeling may improve capture of higher-level authorial style.

8. Conclusion and Future Work

This paper presented a transformer-based system for multiclass authorship identification developed for the AbjadAuthorID shared task. Our approach fine-tunes AraBERT v2 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 directions include longer-context modeling, segment aggregation, 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 Jarrar, 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 Conference of the European Chapter of the Association for Computational Linguistics (EACL 2026)*, Rabat, Morocco, March 2026. 2

[2] Wissam Antoun, Fady Baly, and Hazem Hajj. AraBERT: Transformer-based model for Arabic language understanding. 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 Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France, May 2020. European Language Resource Association. 2

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, 2019. 2

[4] Maël Fabien, Esaú Villatoro-Tello, Petr Motlicek, and Shantipriya Parida. Bertaa: Bert fine-tuning for authorship attribution. In *Proceedings of ICON 2020*, 2020. 2

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

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

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

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