grkurdi at AbjadAuthorID Shared Task: Arabic Author Attribution Using Transformer-Based Models

Ghader Reda Kurdi
Department of Data Science, College of Computing,
Umm Al-Qura University, Makkah, Saudi Arabia
grkurdi@uqu.edu.sa

Abstract

This paper describes the author's participation in the Arabic track of the AbjadAuthorID shared task, which focuses on multiclass authorship attribution using transformer-based models. The task involves identifying the author of a given text excerpt drawn from diverse genres and historical periods, posing significant challenges due to stylistic variation and linguistic richness. Experimental results demonstrate strong performance, with an ensemble of MARBERTv2 and ARBERTv2 achieving an accuracy of 92% and a macro-averaged F1-score of 89%, ranking second on the leaderboard, and highlighting the effectiveness of the proposed approach for Arabic authorship identification.

1 Introduction

Authorship identification is a well-established problem in Natural Language Processing (NLP), concerned with determining or verifying the author of a given text. It has a wide range of applications in multiple areas, including digital humanities, literary analysis, and plagiarism detection. While substantial progress has been achieved for English and other high-resource languages, authorship identification for languages using the Arabic (Abjad) script remains comparatively underexplored.

The AbjadAuthorID shared task (Abudalfa et al., 2026), introduced as part of the 2nd Workshop on NLP for Languages Using Arabic Script (AbjadNLP 2026), addresses this gap by proposing a multiclass authorship attribution challenge focused on literary texts written in languages that use Arabic script, including Modern Standard Arabic (MSA), Urdu, and Kurdish, with each language treated as a separate challenge. The task requires identifying the author of a given text excerpt drawn from diverse genres and historical periods. It involves discriminating among multiple candidate authors, thereby substantially increasing both its methodological complexity and practical relevance.

The author participated in the Arabic track of the shared task, and this paper presents the proposed approach, experimental setup, and evaluation results. Two transformer-based models, ARBERTv2 and MARBERTv2, were trained and evaluated independently, followed by an ensemble approach that combines their predictions to improve overall performance. The proposed ensemble achieved strong results, ranking second on the official leaderboard, and demonstrated good generalization from validation to test data. To support reproducibility and future research, the implementation code is publicly available at: https://colab.research.google.com/drive/1SalZoth2IxTqR1NmeRDDjeuy-i8yqdeR?usp=sharing

2 Background

2.1 Task and Data Description

The task is formulated as a multiclass authorship attribution problem, where the objective is to identify the author of a given text excerpt. The input is a text written in the style of a specific author. The output is the predicted author name in Arabic (using exactly the same author names as those provided in the dataset). Formally, given an input text x, the model predicts a label $y \in A$, where A is the set of candidate authors. To evaluate performance, participants are required to submit a ZIP file containing a single UTF-8 encoded CSV file. The CSV file should include two columns: id, corresponding to the sample identifier, and label, representing the predicted author name.

This work focuses on the Arabic track of the shared task. The dataset is available on Codabench, with a total size of 47,692 instances, distributed across three files:

- `AuthorshipClassficiationTrain.xlsx`, containing 35,122 instances;

Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script (AbjadNLP 2026), pages 520–524
March 28, 2026. ©2026 Association for Computational Linguistics

Table 1: Author distribution across training and validation sets

Author	Train	Train (%)	Val	Val (%)
Hassan Hanafi	3,744	10.66	548	13.18
Ahmed Amin	2,897	8.25	246	5.92
Jurji Zaydan	2,768	7.88	327	7.87
Robert Barr	2,682	7.64	82	1.97
Mohamed Hussein Heikal	2,627	7.48	260	6.25
Taha Hussein	2,376	6.76	255	6.13
Abbas Mahmoud Al-Aqqad	1,829	5.21	267	6.42
Fouad Zakaria	1,773	5.05	125	3.01
Naguib Mahfouz	1,634	4.65	327	7.87
Ameen Rihani	1,567	4.46	142	3.42
Abdel-Ghaffar Mikkawi	1,520	4.33	396	9.53
Gustave Le Bon	1,515	4.31	150	3.61
Nawal El Saadawi	1,415	4.03	295	7.10
William Shakespeare	1,236	3.52	238	5.73
Youssef Idris	1,140	3.25	120	2.89
Salama Moussa	984	2.80	119	2.86
Ahmed Taymour Pasha	815	2.32	57	1.37
Tharwat Abaza	757	2.16	90	2.17
Gibran Khalil Gibran	750	2.14	30	0.72
Ahmed Shawqi	679	1.93	58	1.40
Kamel Kilani	414	1.18	25	0.60

- `AuthorshipClassficiationVal.xlsx`, containing 4,157 instances;

Each file contains three columns:

- **id**: a unique identifier for each text excerpt,
- **text_in_author_style**: the input text written in the style of a specific author,
- **author**: the ground-truth author name corresponding to text excerpts.

After the development phase ended, an additional file, `PublicDataFinalPhaseTask2.xlsx`, containing 8,413 instances, was released for the final evaluation phase. This file contains only two columns, **id** and **text_in_author_style**, with the author labels withheld to allow blind evaluation.

According to the organizers, the corpus comprises texts from 21 classical and modern authors, collected from 10 publicly accessible books per author. Each book was automatically segmented into semantically coherent paragraphs using the Natural Language Toolkit (NLTK). The distribution of authors in the development and validation sets is shown in Table 1 . Additional details on the corpus and its development are provided in (Abudalfa et al., 2025).

2.2 Related Work

Research on Arabic authorship identification remains limited; a survey of studies published between 2010 and 2020 (Alqahtani and Dohler, 2023) identified only 20 studies focused on Arabic authorship attribution. Most of these studies relied on shallow machine learning models and manually engineered features. The survey concluded that the reported results vary significantly depending on the selected feature sets and the datasets used, with the effective features differing between the classical Arabic, Modern Standard Arabic, and Colloquial Arabic texts. The limited availability of publicly accessible datasets was identified as a major factor contributing to the scarcity of research in this area. The author emphasized that a key research priority should be the development of diverse, well-curated, and openly accessible datasets, enabling further research and more comparable results in Arabic authorship attribution. This issue was addressed in (Abudalfa et al., 2025) through the development of a large-scale Arabic authorship attribution dataset and the organization of a shared task to support experimentation and comparable evaluation. The first challenge was introduced through the AraGenEval Shared Task, hosted at the ArabicNLP 2025 conference, where the highest performing system (Helmy et al., 2025) achieved a macro-F1 score of 90% using an ensemble of four transformer-based models with final predictions computed using soft-voting over model outputs. This effort was subsequently extended through the AbjadAuthorID shared task.

3 System Overview

Several Arabic transformer-based models, including XLM-R, CAMeLBERT, MARBERTv2, and ARBERTv2, were initially evaluated. Based on early experimental results, MARBERTv2 and ARBERTv2 were selected and fine-tuned on the shared task data. In addition, a lightweight late-fusion ensemble is constructed by combining the prediction probabilities of the individual models. The ensemble operates by averaging the predicted class probabilities produced by each model at inference time and selecting the class with the highest aggregated confidence score, where w_1 and w_2 denote the ensemble weights (both set to 0.5).

4 Experimental Setup

4.1 Data Splitting and Evaluation

During the development phase, we followed the data split provided by the organizers, using the training set to train the model and select hyperparameters and the number of training epochs, while the validation set was used exclusively to measure performance. For the final evaluation phase, we retrained the selected models by combining the training and validation sets, keeping all hyperparameters unchanged, and submitted the resulting predictions for evaluation on the test set.

4.2 Preprocessing

Preprocessing was limited to model-specific tokenization using MARBERTv2 or ARBERTv2, with input sequences truncated to 384 tokens and dynamically padded during batching. No additional preprocessing was applied.

4.3 Parameter Settings

MARBERTv2 and ARBERTv2 were fine-tuned using the Hugging Face Trainer framework under largely similar experimental settings, as reported in Table 2.

4.4 Evaluation metrics

Following the organizers' guidelines, performance is primarily evaluated using the macro-F1 score, with accuracy reported as a secondary metric.

Let N denote the total number of test instances, C the set of classes, y_i the true label of instance i, $\hat{y}_i$ the predicted label, and $I(\cdot)$ the indicator function.

$$Accuracy = \frac{1}{N}\sum_{i=1}^{N} I(\hat{y}_i = y_i)$$

For each class $c \in C$, let TP_c, FP_c, and FN_c denote the numbers of true positives, false positives, and false negatives, respectively. Precision and recall are defined as:

$$Precision_c = \frac{TP_c}{TP_c + FP_c},$$

$$Recall_c = \frac{TP_c}{TP_c + FN_c}$$

The class-wise F1-score is:

$$F1_c = \frac{2 \cdot Precision_c \cdot Recall_c}{Precision_c + Recall_c}$$

The macro-F1 score is computed as the unweighted average over all classes:

$$Macro - F1 = \frac{1}{|C|}\sum_{c \in C} F1_c$$

5 Results

According to the official evaluation results released by the organizers, the proposed approach ranked second among the eight participating teams on the leaderboard. The highest performance was achieved by the MARBERTv2 and ARBERTv2 ensemble, attaining a macro-F1 score of 92.44% and an accuracy of 88.97% on the test set. The top-ranked system obtained a macro-F1 score of 93.21% and an accuracy of 96.34, indicating comparable performance under the shared-task evaluation protocol. The results obtained by ARBERTv2 are very close to those of the ensemble, indicating that ARBERTv2 alone provides strong and competitive performance. These results confirm the effectiveness of transformer-based and ensemble modeling strategies for Arabic authorship attribution. Comparing the results on the validation and test sets (Table 3) indicates good generalization to unseen data, with improved performance observed on the test set.

Table 2: Final hyperparameter configuration used for retraining on the combined training and validation sets.

Parameter	MARBERTv2	ARBERTv2
Base model	UBC-NLP/MARBERTv2	UBC-NLP/ARBERTv2
Max sequence length	384	384
Tokenizer	Fast tokenizer	Fast tokenizer
Padding	Dynamic (DataCollatorWithPadding)	Dynamic (DataCollatorWithPadding)
Epochs	5	6
Learning rate	$1e^{-5}$	$2e^{-5}$
Batch size	4	4
Weight decay	0.01	0.01
Warm-up ratio	0.06	0.06
Random seed	42	42
Precision	FP16 (when available)	FP16 (when available)
Evaluation during training	None	None
Checkpointing	Every epoch	Every epoch
Prediction	Softmax + Argmax	Softmax + Argmax

In contrast to prior work relying on large multi-model ensembles (Helmy et al., 2025), the proposed approach achieves competitive performance using a substantially simpler ensemble, highlighting a favorable trade-off between performance and model complexity.

Split	Model	Macro-F1	Accuracy
Valid.	MARBERTv2	81.94	88.62
	ARBERTv2	85.94	90.11
	Ensemble	86.37	91.10
Test	MARBERTv2	83.96	88.41
	ARBERTv2	87.53	91.35
	Ensemble	88.97	92.44

Table 3: Performance of MARBERTv2, ARBERTv2, and their ensemble on the validation (valid.) and test sets.

5.1 Error Analysis

Certain authors, such as Tharwat Abaza, Ahmed Shawqi, and Mohamed Hussein Heikal, proved challenging for all models. Detailed results are presented in Table 4 in the Appendix, which reports the per-author F1-scores on the validation set for MARBERTv2, ARBERTv2, and their ensemble.

An error analysis of MARBERTv2 reveals that the most frequent confusion occurs between Tharwat Abaza and Ahmed Shawqi (49 instances), followed by Mohamed Hussein Heikal misclassified as Tharwat Abaza (41 instances). Additional prominent confusions include Tharwat Abaza misclassified as Naguib Mahfouz (20 instances), as well as both Mohamed Hussein Heikal and Abbas Mahmoud Al-Aqqad misclassified as Ahmed Amin (18 instances each).

In contrast, ARBERTv2 exhibits a slightly different confusion pattern. The most frequent error involves Tharwat Abaza and Mohamed Hussein Heikal (39 instances), followed by misclassifications of Abdel-Ghaffar Mikkawi as Fouad Zakaria (24 instances) and Tharwat Abaza as Ahmed Shawqi (22 instances). Additional notable confusions include William Shakespeare misclassified as Fouad Zakaria (20 instances), Mohamed Hussein Heikal misclassified as Ahmed Amin (19 instances), and Abbas Mahmoud Al-Aqqad misclassified as Ahmed Amin (15 instances).

6 Conclusion

This study investigates Arabic authorship attribution using pretrained transformer models, namely MARBERTv2 and ARBERTv2, along with a simple late-fusion ensemble strategy. The ensemble, based on simple probability averaging, showed strong performance on the AbjadAuthorID shared task, demonstrating good generalization from validation to test data and achieving second place. However, the experiments are conducted exclusively on the AbjadAuthorID dataset, which represents a specific set of Arabic authors and writing styles. As a result, the findings may not directly generalize to other Arabic corpora with different genres or domains. Further evaluation on additional datasets would be required

References

Shadi Abudalfa, Saad Ezzini, Ahmed Abdelali, Hamza Alami, Abdessamad Benlahbib, Salmane Chafik, Mo El-Haj, Abdelkader El Mahdaouy, Mustafa Jarrar, Salima Lamsiyah, and Hamzah Luqman. 2025. The AraGenEval shared task on Arabic authorship style transfer and AI generated text detection. In *Proceedings of The Third Arabic Natural Language Processing Conference at EMNLP 2025*, Suzhou, China. Association for Computational Linguistics.

Shadi Abudalfa, Saad Ezzini, Ahmed Abdelali, Mustafa Jarrar, Mo El-Haj, Nadir Durrani, Hassan Sajjad, Farah Adeeba, and Sina Ahmadi. 2026. 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.

Fatimah Alqahtani and Mischa Dohler. 2023. Survey of authorship identification tasks on arabic texts. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(4):1–24.

Muhammad Helmy, Batool Najeh Balah, Ahmed Mohamed Sallam, and Ammar Sherif. 2025. Sebaweh at arageneval shared task: Berense-bert based ensembler for arabic authorship identification. In *Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks*, pages 59–64.

A Detailed Validation Results

Table 4: F1_score on validation set. H.H.: Hassan Hanafi, A.A.: Ahmed Amin, J.Z.: Jurji Zaydan, R.B.: Robert Barr, M.H.H.: Mohamed Hussein Heikal, T.H.: Taha Hussein, A.M.A.: Abbas Mahmoud Al-Aqqad, F.Z.: Fouad Zakaria, N.M.: Naguib Mahfouz, A.R.: Ameen Rihani, A.G.M.: Abdel-Ghaffar Mikkawi, G.L.B.: Gustave Le Bon, N.E.S.: Nawal El Saadawi, W.S.: William Shakespeare, Y.I.: Youssef Idris, S.M.: Salama Moussa, A.T.P.: Ahmed Taymour Pasha, T.A.: Tharwat Abaza, G.K.G.: Gibran Khalil Gibran, A.S.: Ahmed Shawqi, K.K.: Kamel Kilani.

Author	**MARBERT**	**ARBERT**	**Ensemble**
H.H.	98.37	98.55	98.91
A.A.	85.01	84.77	86.12
J.Z.	94.27	93.19	95.27
R.B.	93.02	96.97	96.39
M.H.H.	69.03	74.95	78.48
T.H.	94.32	95.02	94.96
A.M.A.	88.93	89.44	90.37
F.Z.	85.12	82.55	89.86
N.M.	93.70	97.58	97.26
A.R.	82.44	85.61	87.46
A.G.M.	93.54	93.40	94.01
G.L.B.	93.04	91.37	93.85
N.E.S.	97.07	96.14	96.40
W.S.	87.31	88.30	87.95
Y.I.	92.19	88.97	92.19
S.M.	98.35	92.25	93.33
A.T.P.	84.38	84.13	72.06
T.A.	8.64	24.14	20.95
G.K.G.	73.33	91.80	98.36
A.S.	59.39	66.67	70.31
K.K.	49.18	88.89	79.37