Uslūb at AbjadAuthorID Shared Task: A Comparative Analysis of Traditional Machine Learning and Transformer-Based Models for Authorship Attribution in Arabic and Urdu

Shahad Alsuhaibani
Department of Computer Science
King Saud University
Riyadh, Saudi Arabia

Mohamed Alkaoud
Department of Computer Science
King Saud University
Riyadh, Saudi Arabia
malkaoud@ksu.edu.sa

Abstract

Authorship attribution is a critical task in natural language processing with applications ranging from forensic linguistics to plagiarism detection. While well-studied in high-resource languages, it remains challenging for low-resource languages like Arabic and Urdu. In this paper, we present our participation in the *AbjadNLP* shared task, where we systematically evaluate three distinct approaches: traditional machine learning using SVM with TF-IDF features, fine-tuned transformer-based models (AraBERT), and LLMs. We demonstrate that while fine-tuned AraBERT excels in Arabic, traditional lexical models (SVM) prove more robust for Urdu, outperforming both BERT-based and LLM approaches. We also show that few-shot prompting with LLMs, when operated as a reranker over top candidates, significantly outperforms zero-shot baselines. Our final systems achieved competitive performance, ranking 6th and 1st in the Arabic and Urdu tasks respectively.

1 Introduction

Authorship attribution is a fundamental problem in natural language processing, aiming to identify the author of a text based on stylistic, lexical, and contextual patterns. It plays an important role in applications such as literary analysis, plagiarism detection, and forensic linguistics. While the task has been extensively studied for high-resource languages (Stamatatos, 2009; Kestemont, 2014; Hung et al., 2023; Gorovaia et al., 2024; Hu et al., 2024), it remains challenging for morphologically rich and low-resource languages, where stylistic variation, data imbalance, and limited annotated corpora complicate model generalization.

In this work, we address authorship attribution for Arabic and Urdu within the *AbjadNLP* (Abudalfa et al., 2026, 2025) shared task. Although both languages are written in the Arabic script, they differ substantially in morphology, syntax, and vocabulary, reflecting their distinct linguistic origins; Arabic being a Semitic language and Urdu being an Indo-Aryan language. This makes them an interesting case for cross-language comparison. We investigate a range of approaches, including traditional machine learning methods, pretrained BERT-based language models, and large language models, to assess their effectiveness across the two languages in the authorship attribution task. Our work evaluates how well these approaches capture author-specific signals in both Arabic and Urdu, and analyzes the extent to which methods effective in one language generalize to the other. Based on our experimental results, our system ranked 6th in Arabic and achieved 1st place in Urdu on the shared task leaderboard.

2 Background

Authorship attribution has been widely explored across languages. Traditional approaches use features like word and character n-grams or Term Frequency–Inverse Document Frequency (TF-IDF) with baseline classifiers, achieving strong performance in high-resource settings (Stamatatos, 2009; Kestemont, 2014). More recently, BERT-based pretrained language models such as AraBERT (Antoun et al., 2020) and AraELECTRA (Elmadany et al., 2021) have become dominant in Arabic authorship attribution, effectively capturing rich morphological and semantic cues and shown strong performance (AlZahrani and Al-Yahya, 2023). Additionally, large language models have been investigated for authorship tasks (Hung et al., 2023; Hu et al., 2024), demonstrating promising performance, especially when limited examples of candidate authors are provided. In this work we evaluate the effectiveness of these approaches for authorship attribution in Arabic and Urdu.

Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script (AbjadNLP 2026), pages 515–519

3 Dataset Details

The dataset used for training was provided by the *AbjadNLP* shared task organizers for author attribution tasks in both Arabic and Urdu. The Arabic dataset consists of texts from 21 authors, each represented by 10 publicly available books, while the Urdu dataset includes texts from 10 authors. For both languages, the texts were segmented into semantically coherent paragraphs and organized into training, validation, and test splits. Both datasets are imbalanced across authors, with varying numbers of samples per author. Figures 1 and 2 show the number of training and validation samples per author for Arabic and Urdu. As noticed, there's a noticeable imbalance among the classes.

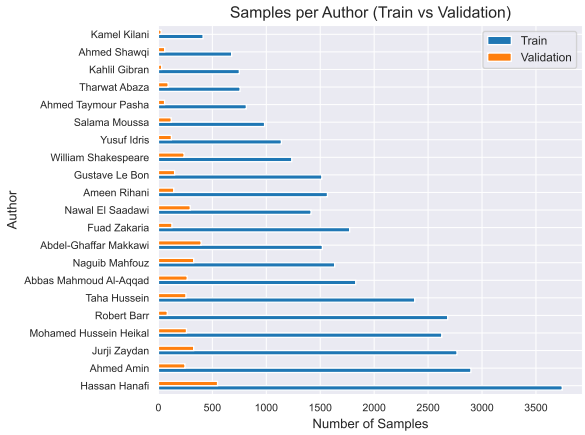

Figure 1: Samples per author in Train and Validation sets of the Arabic dataset.

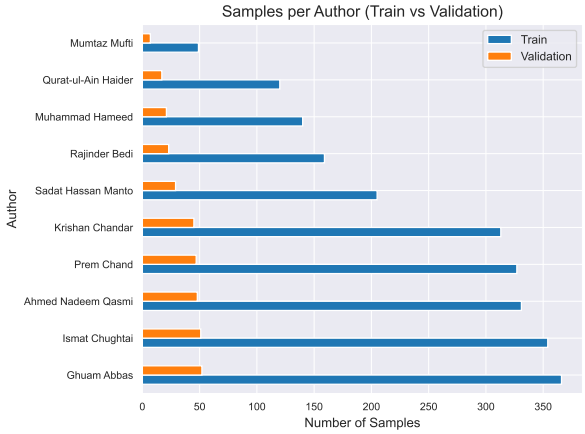

Figure 2: Samples per author in Train and Validation sets of the Urdu dataset.

4 System Overview

Our system for the *AbjadNLP* shared task adopts different approaches to authorship attribution, recognizing that authorial style manifests across lexical, syntactic, and semantic dimensions. First, we employ a Support Vector Machine (SVM) with TF-IDF (Sparck Jones, 1972) features; while traditional, this remains a potent method for capturing the lexical and syntactic stylometry often sufficient for attribution, independent of semantic content. Second, we fine-tune AraBERT to model deeper linguistic context. This decision is grounded in recent benchmarks, such as LAraBench (Abdelali et al., 2024), which demonstrate that fine-tuned models frequently outperform LLMs in Arabic tasks. Finally, despite these benchmarks, we explore the capabilities of generative LLMs, aiming to determine if their documented success in English authorship attribution extends effectively to Arabic and Urdu.

4.1 Baseline: SVM with TF-IDF

For the traditional machine learning baseline, we employ an SVM classifier trained on TF-IDF representations of the input texts. Texts are vectorized using TF-IDF (Sparck Jones, 1972) features with an n-gram range of unigrams and bigrams, enabling the model to capture both individual lexical units and short compound patterns that are indicative of authorial style. The resulting TF-IDF feature matrix is used to train an SVM with a linear kernel, which is well suited for high-dimensional sparse feature spaces.

4.2 Fine-tuned BERT

We utilize one of the most popular Arabic pretrained language models, `bert-base-arabertv02`, which is a 12-layer bidirectional Transformer encoder based on BERT (Devlin et al., 2019). AraBERT is pretrained on large-scale Arabic corpora including news, Wikipedia, and social media using the Masked Language Modeling (MLM) objective. The model is fine-tuned for multi-class author classification and applied to both Arabic and Urdu datasets to examine the extent to which an Arabic-pretrained transformer can generalize to a language with the same script. Fine-tuning is performed for 7 epochs for Arabic and 3 epochs for Urdu due to the smaller dataset size. During training, a classification head is learned on top of the contextualized document representations to predict the target author labels, using a learning

rate of 5×10^{-6} and a batch size of 8.

4.3 Large Language Models

For the large language model experiments, we employ `gpt-5-chat` for authorship attribution using both two settings.

In the first setting, the model is tasked with predicting the author without any examples, but with the complete list of candidate authors provided. This configuration evaluates the model's capacity to leverage its vast pre-trained knowledge for author discrimination. We will refer to this setting moving forward as zero-shot.

For the second setting, we implement a constrained few-shot strategy. Unlike standard few-shot classification where the model must select from the entire author pool, we effectively treat the LLM as a reranker to refine the predictions.

This process operates in two stages. First, we drastically reduce the search space by using the best-performing model for each language to retrieve only the top-3 most probable authors. Second, we prompt the LLM to select the correct author from this narrowed subset. This approach serves a dual purpose: it simplifies the classification task (reducing the decision boundary to 3 classes) and allows us to provide targeted few-shot examples, two per candidate, without exceeding the model's effective context window or inducing the 'lost-in-the-middle' phenomenon (Du et al., 2025; Liu et al., 2024). Attempting to include few-shot examples for the entire label space (e.g., 21 authors for Arabic) would necessitate concatenating dozens of documents into a single prompt. Such extended contexts are known to degrade LLM capabilities. By narrowing the scope to just three candidates, we maintain a compact and high-density context. We will refer to this setting moving forward as few-shot.

For both Arabic and Urdu texts, the LLM instructions were given in English, as prior studies have shown that LLMs can perform better with English prompts in non-English tasks, even when processing non-English text (Dey et al., 2024; Lai et al., 2023; Alkaoud, 2024). Figure 3 illustrates the prompt template employed in the few-shot LLM experiments.

```
Instructions: You are an expert in authorship
attribution. Decide which author most likely
wrote the text. Provide your answer only as
the author's name.

Candidate Author Samples:
<author A>: <text 1>; <text 2>
<author B>: <text 1>; <text 2>
<author C>: <text 1>; <text 2>

Text to Classify: <author text to classify>
```

Figure 3: Prompt used for few-shot LLM experiments.

5 Experimental Setup

Experiments using SVM with TF-IDF features were executed on a CPU. AraBERT was fine-tuned on Google Colab with an NVIDIA A100 GPU, with a maximum sequence length of 512 tokens and a batch size of 8. Inputs larger than 512 tokens were truncated. All the LLM experiments using GPT-5-chat were performed via the OpenAI API.

5.1 Evaluation Metrics

Following the shared task guidelines, we report Macro F1 as the primary metric to account for class imbalance, and accuracy as a secondary metric. Both are computed on the validation set.

6 Results and Discussion

Table 1 summarizes the performance of all models on the validation set across Arabic and Urdu. For Arabic, AraBERT achieved the best performance, with an accuracy of 0.878 and a Macro F1-score of 0.8195. TF-IDF features with SVM achieved comparable performance, demonstrating that lexical patterns are highly informative for Arabic.

In contrast, zero-shot GPT-5 prompting performed poorly (accuracy 0.137, macro F1 0.108), showing that the model struggled to discriminate authors. To provide context, the top-3 candidate authors were generated using the best-performing model. Incorporating two examples per candidate in a few-shot prompt substantially improved GPT-5 performance (accuracy 0.870, macro F1 0.806), although AraBERT remained the most effective approach for Arabic authorship attribution.

Figure 4 shows the normalized confusion matrix for AraBERT on the Arabic validation set. It shows that the model correctly identifies most authors with high accuracy. Notably, performance is lower for some minor authors, including Kilani, Abaza, and Shawqi, who are frequently confused with other authors. This indicates that, despite its overall strength, AraBERT still struggles with authors represented by fewer samples.

For Urdu, AraBERT, struggled in this cross-lingual setting, achieving only 0.229 accuracy and 0.171 macro F1. While TF-IDF/SVM model per-

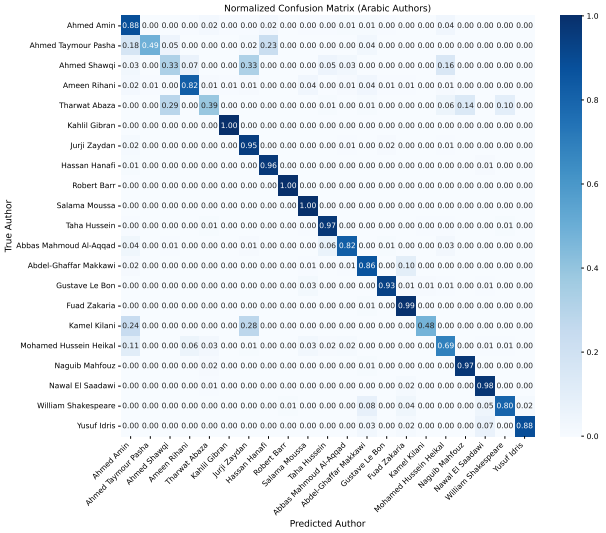

Figure 4: Normalized confusion matrix of Arabic authors on the validation set using AraBERT fine-tuned.

Figure 5: Normalized confusion matrix of Urdu authors on the validation set using SVM + TF-IDF.

Model	Arabic		Urdu	
	Acc.	F1	Acc.	F1
SVM + TF-IDF	89.25	79.12	45.29	**45.17**
AraBERT	87.75	**81.95**	22.94	17.11
GPT-5 (zero-shot)	13.70	10.75	17.65	16.50
GPT-5 (few-shot)	87.02	80.55	33.82	38.14

Table 1: Performance of the different models on Arabic and Urdu validation sets.

formed better, it is still much worse compared to Arabic. The top-3 SVM predictions reached 0.674 accuracy, suggesting that traditional lexical representations can still provide some useful signals in Urdu. GPT-5 zero-shot achieved very low performance (0.177 accuracy, 0.165 macro F1). While few-shot prompting substantially improved performance for Arabic, it was less effective for Urdu (0.338 accuracy, 0.381 macro F1), indicating that LLM effectiveness varies across the two languages. To isolate the reasoning capability of the LLM from the retrieval limitations, we calculate the normalized accuracy based on the top-3 accuricies for each language as shown in Table 2. In Arabic, the normalized accuracy is 0.914, indicating that the LLM is highly effective at distinguishing between plausible candidates when the ground truth is present. In contrast, the Urdu normalized accuracy is 0.502, suggesting that the model struggles to captured to Arabic. Figure 5 shows the confusion matrix for the Urdu SVM model. Unlike the Arabic results, the Urdu matrix shows significant inter-author confusion and a less defined diagonal.

Language	Model	Accuracy (%)
Arabic	AraBERT	95.26
Urdu	SVM + TF-IDF	67.35

Table 2: Top-3 predictions accuracy for Arabic and Urdu.

Based on the validation set results, we selected the best-performing model for each language and submitted it for evaluation on the blind test set. For Arabic, the AraBERT-based system achieved an accuracy of 0.869 and an F1-score of 0.836. For Urdu, the SVM model using TF-IDF features achieved an accuracy of 0.355 and an F1-score of 0.395.

7 Conclusion and Future Work

We presented three different approaches for authorship attribution in Arabic and Urdu in the *AbjadNLP* shared task, comparing traditional machine learning methods, pretrained BERT-based models, and LLMs. Our results show that AraBERT is effective for Arabic, while TF-IDF with SVM performs better for Urdu and coming close to the performance of AraBERT in Arabic. Utilizing LLMs did not lead to substantial performance gains in this setting.

Our final systems achieved competitive leaderboard results, ranking 6th in Arabic and 1st in Urdu. For future work, we plan to investigate hybrid models that combine SVM-based stylistic features with BERT-based semantic representations, aiming to jointly capture lexical style and deeper contextual information for improved authorship attribution.

References

Ahmed Abdelali, Hamdy Mubarak, Shammur Chowdhury, Maram Hasanain, Basel Mousi, Sabri Boughorbel, Samir Abdaljalil, Yassine El Kheir, Daniel Izham, Fahim Dalvi, Majd Hawasly, Nizi Nazar, Youssef Elshahawy, Ahmed Ali, Nadir Durrani, Natasa Milic-Frayling, and Firoj Alam. 2024. LAraBench: Benchmarking Arabic AI with large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 487–520, St. Julian's, Malta. Association for Computational Linguistics.

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.

Mohamed Alkaoud. 2024. A bilingual benchmark for evaluating large language models. *PeerJ Computer Science*, 10:e1893.

Fetoun Mansour AlZahrani and Maha Al-Yahya. 2023. A transformer-based approach to authorship attribution in classical arabic texts. *Applied Sciences*, 13(12):7255.

Wissam Antoun and 1 others. 2020. Arabert: Transformer-based model for arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools*.

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

Krishno Dey, Prerona Tarannum, Md. Arid Hasan, Imran Razzak, and Usman Naseem. 2024. Better to ask in english: Evaluation of large language models on english, low-resource and cross-lingual settings. *arXiv preprint*.

Yufeng Du, Minyang Tian, Srikanth Ronanki, Subendhu Rongali, Sravan Bodapati, Aram Galstyan, Azton Wells, Roy Schwartz, Eliu A. Huerta, and Hao Peng. 2025. Context length alone hurts llm performance despite perfect retrieval. In *Findings of EMNLP 2025*.

Ahmed Elmadany and 1 others. 2021. Araelectra: Pre-training text discriminators for arabic language understanding. In *Proceedings of the 3rd Arabic Natural Language Processing Workshop*.

Svetlana Gorovaia, Gleb Schmidt, and Ivan P. Yamshchikov. 2024. Sui generis: Large language models for authorship attribution and verification in latin. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 398–412. Association for Computational Linguistics.

Zhengmian Hu, Tong Zheng, and Heng Huang. 2024. A bayesian approach to harnessing the power of llms in authorship attribution. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13216–13227. Association for Computational Linguistics.

Chia-Yu Hung, Zhiqiang Hu, Yujia Hu, and Roy Lee. 2023. Who wrote it and why? prompting large-language models for authorship verification. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14078–14084. Association for Computational Linguistics.

Mike Kestemont. 2014. Function words in authorship attribution: From black magic to theory? In *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLfL)*, pages 59–66. ACL.

Viet Dac Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189, Singapore. Association for Computational Linguistics.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556.