R-R at AbjadAuthorID Shared Task: A Fine-Tuned Approach for Kurdish Authorship Identification

Rania Azad M. San Ahmed
Computer Networks Department
Sulaimani Polytechnic University
Sulaimani, Iraq
rania.azad@spu.edu.iq

Rebwar M. Nabi
Deanery Office
Sulaimani Polytechnic University
Kurdistan Technical Institute
Sulaimani, Iraq
rebwar.nabi@kti.edu.iq

Abstract

Authorship identification is a fundamental task in natural language processing and computational stylistics. Despite significant advancements in high-resource languages, low-resource languagesparticularly those utilizing non-Latin scriptsremain largely underexplored, leaving a critical gap in resources and benchmarks for this linguistically distinct, low-resource language. Addressing this oversight, this paper presents Task 3 of AbjadNLP 2026, the first shared task dedicated to authorship identification for Kurdish. The task introduces a newly constructed dataset designed to capture the unique phonological and orthographic features of Sorani Kurdish and formulate the task as a closed-set multiclass classification problem. To establish a robust baseline, we fine-tune the pretrained XLM-RoBERTa model to capture authorial, stylistic patterns. Experimental results on the test set demonstrate the efficacy of transformer-based representations for this domain, achieving an accuracy of approximately 75%.

1 Introduction

Authorship identification is a fundamental task in Natural Language Processing (NLP) and computational linguistics, aiming to determine the author of a given text based on stylistic and linguistic cues. It has wide-ranging applications in digital humanities, literary analysis, forensic linguistics, and plagiarism detection. While the problem has been extensively studied for English and other Latin-script languages, substantially less attention has been devoted to languages written in the Arabic (Abjad) script, particularly in low-resource and morphologically rich settings.

Recent shared tasks have contributed to advancing research in this direction. The AraGenEval 2025 shared task on Arabic authorship attribution (Abudalfa et al., 2025) introduced a benchmark for evaluating computational approaches to identifying authors from collections of Arabic literary texts, highlighting both the challenges and opportunities of authorship modeling in Arabic. These efforts demonstrate the importance of standardized evaluation frameworks for Arabic-script languages but also expose a notable gap in coverage for other languages that share the same script yet differ substantially in linguistic structure.

This year, AbjadNLP 2026 extended the task of authorship identification research beyond Arabic by adding Urdu and Kurdish (Sorani) (Abudalfa et al., 2026). The Kurdish language is from the Indo-Iranian branch language family and shares close linguistic similarities with Persian and Arabic. It is spoken by an estimated 30 to 40 million people across Iraq, Iran, Turkey, Armenia, and Syria. The language is characterized by a diverse dialectal landscape, with Kurmanji (Northern Kurdish) and Sorani (Central Kurdish) being the two most widely spoken varieties. Sorani, in particular, employs the Perso-Arabic script, consisting of 36 characters (33 consonants and 3 vowels), and is written from right to left. Kurdish remains underrepresented in both academic research and technological development compared to more widely studied languages (San Ahmed and Saeed, 2025).

This work contributes a baseline transformer-based approach for authorship identification in Kurdish by fine-tuning XLM-RoBERTa, a multilingual model pretrained on diverse languages using the Arabic script. The proposed system performs multiclass classification of text excerpts by author and serves as an initial benchmark for Kurdish authorship identification within the AbjadNLP shared-task framework.

2 Related Work

Kurdish remains a low-resource language with limited annotated datasets, and research in Kur-

Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script (AbjadNLP 2026), pages 525–529

dish natural language processing is still emerging. Nevertheless, recent years have witnessed growing interest in Kurdish NLP across several tasks such sentiment analysis for Kurdish texts, focusing on polarity classification (Badawi et al., 2025a) (San Ahmed and Saeed, 2025) (Karim and Abdullah, 2025) (Badawi, 2023) and text classification (Badawi, 2024) (Badawi et al., 2025b) using both traditional machine learning methods and deep learning models. Furthermore, several studies have been focused on creating resources for different tasks , hope detection (Badawi, 2025), fake news detection (San Ahmed et al., 2021), and Named Entity Recognition (NER) (Abdalla et al., 2025)(Wahid and Nabi, 2025), stance detection (Rostam and Nabi, 2025) and sarcasm detection (Aghajan and Nabi, 2025). Despite these advances, authorship identification for Kurdish has received no attention, and no prior shared task has explicitly addressed this problem.

On the other hand, recent work on the Arabic language has explored different approaches, ranging from traditional machine learning approaches to large language models. The top-performing system in shared task 2025, Sebaweh (Helmy et al., 2025), employed four fine-tuned transformer-based modelsAraBERT, CAMELBERT, Arabic XLM-RoBERTa, and GATE-AraBERTdemonstrating the effectiveness of model diversity and ensembling in capturing authorial style. Similarly, the team, Athership (Samir et al., 2025), adopted an ensemble strategy based on dual-model logit fusion, combining AraBERT and AraELECTRA to enhance classification. Large language models were also explored in the competition. The MISSION team (Alharbi, 2025), which ranked fourth, fine-tuned the ALLaM-7B-Instruct-preview model using prompt engineering techniques, highlighting the potential of instruction-tuned models for authorship attribution tasks. In contrast, several participants demonstrated that competitive performance can still be achieved using lightweight and traditional methods. The team (Sabaa and Sabaa, 2025), ranked eighth, combined word-level and character-level TFIDF features with a logistic regression classifier, underscoring the continued relevance of classical feature-based approaches in authorship identification. Likewise, NLP wizard (Hany, 2025) utilized pre-trained XLM-RoBERTa embeddings as fixed feature extractors, followed by classical classifiers such as LinearSVC.

No	Author(English)	Train	Validation
1	Hazhar	1673	240
2	Ibrahim Ahmed	958	137
3	Ahlam Mansour	952	136
4	Ara Ilikhanzada	868	125
5	Hemn	625	89
6	Aladdine Sajadi	551	81
7	Hasan Kazlaji	428	61
8	Mala Karim Sarda Kosani	241	33
9	Ali Hassaniani	180	26
10	Ahmed Mokhtar Jaf	168	24
11	Jamal Nabaz	161	23
12	Karim Bagui Jaf	155	22
13	Mala Mohamadi Chrostani	63	9
14	Mala Gawra	30	4
15	Ziwar	25	4

Table 1: Author identification dataset statistics by author and data split.

Statistic	Training	Validation
Total samples	7099	1017
Number of authors	16	16
Mean text length (chars)	247.67	224.90
Median text length (chars)	124	104
Largest author (samples)	1673	240
Smallest author (samples)	21	3
Mean text length (words)	43.14	39.06
Median text length (words)	22	18

Table 2: Dataset statistics.

3 Dataset

The dataset used in this study was provided by AbjadNLP 2026 and is publicly available online (https://ezzini.github.io/AbjadAuthorID/). The dataset contains 7,099 training samples and 1,017 validation samples from 16 Kurdish (Sorani) authors. The distribution of samples across authors is naturally imbalanced, with the author Hazhar being the most represented author (1,673 training, 240 validation), followed by Ibrahim Ahmed (958/137). Several authors have fewer than 200 training samples; however, all authors are represented in both splits as shown in Table 1 and Table 2 .

This stratified distribution reflects realistic low-resource literary data and supports a closed-set multiclass authorship identification setting. Figure 1 presents the text distribution of training and validation samples across authors. Figure 2 depicts the author distribution in the training and validation set.

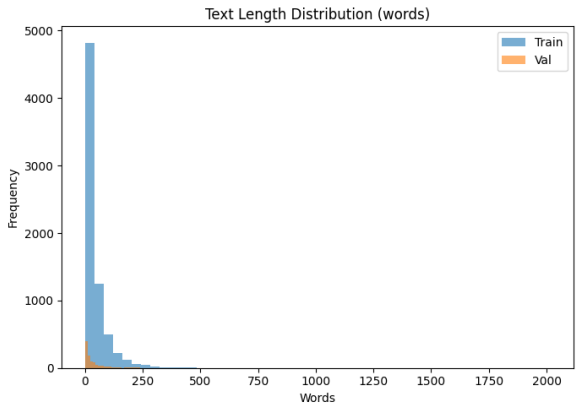

Figure 1: Overall text length distribution in training and validation sets

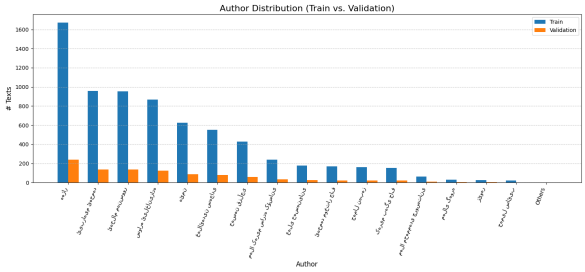

Figure 2: Overall author distribution in training and validation sets

4 Methodology

We formulate the authorship identification task as a closed-set multiclass text classification problem, where each input text segment is assigned to one author from a predefined set of candidates. Our approach is based on fine-tuning the pretrained XLM-RoBERTa model in an end-to-end manner for multiclass authorship identification. A task-specific classification head is added on top of the transformer encoder, and all model parameters are optimized jointly using a weighted cross-entropy loss to account for class imbalance.

XLM-RoBERTa is a multilingual transformer encoder built upon the RoBERTa architecture and pretrained on large-scale multilingual corpora covering more than 100 languages. The model comprises a stack of self-attentionbased transformer encoder layers that produce contextualized representations of input text, enabling effective cross-lingual and stylistic modeling for low-resource languages. The final hidden representation corresponding to the special classification token is fed into a linear classification layer to predict the author label.

Model optimization is performed using the AdamW optimizer with a learning rate of 2e-5 and a weight decay of 0.01. The model is trained for 10 epochs, and performance is evaluated after each epoch on a held-out validation set using Macro-F1 as the primary evaluation metric and accuracy as a secondary metric. The full parameters used in this study are described in Table 3

Component / Parameter	Specification
Model Architecture	
Base Model	XLM-RoBERTa (Pre-trained)
Architecture Type	Transformer Encoder (Multilingual)
Input Representation	Contextualized CLS Token
Classification Head	Linear Layer
Loss Function	Weighted Cross-Entropy
Training Hyperparameters	
Optimizer	AdamW
Learning Rate	2×10^{-5}
Weight Decay	0.01
Training Epochs	10
Batch Strategy	End-to-end Fine-tuning
Evaluation	
Primary Metric	Macro-F1
Secondary Metric	Accuracy

Table 3: Summary of Model Architecture and Hyperparameter Configuration

5 Experimental Setup and Results Analysis

All experiments were conducted within a high-performance computational environment provided by Google Colab Pro+, utilizing a dedicated NVIDIA H100 GPU to ensure efficient training throughput and memory management for the transformer architecture. Table 4 summarizes the quantitative performance of the proposed system on the held-out validation set.

The fine-tuned XLM-RoBERTa model yields a top-1 Accuracy of 75% and a Macro-F1 score of 60%. These results substantiate the efficacy of multilingual transformers in capturing distinct authorial signatures within Sorani Kurdish, a low-resource language with complex morphology. The accuracy metric indicates that, in nearly 75% of cases, the model correctly attributes the text to its true author, establishing a strong baseline for this novel dataset. However, a critical examination of the disparity between accuracy (75%) and macro-F1 (60%) reveals important insights regarding the dataset's distribution. Even though accuracy reflects global correctness, it can be heavily influenced by majority classes. The lower Macro-F1 scorewhich calculates the harmonic

Output	Accuracy	Macro-F1
Validation Set	0.7198	0.6001
Test Set	0.75062	0.59643

Table 4: Model performance on the validation and test sets.

mean of precision and recall for each author independently before averagingsuggests that the models performance is somewhat non-uniform across candidates. This discrepancy implies that while the model excels at identifying authors with prolific writing samples (dominant classes), it faces greater challenges with under-represented authors. Despite this, a Macro-F1 score of 60% in a multiclass setting demonstrates that the weighted cross-entropy loss successfully mitigated the most severe effects of class imbalance, preventing the model from collapsing into a majority-class baseline. Overall, these findings confirm that cross-lingual transfer learning via XLM-RoBERTa is a viable strategy for Kurdish authorship identification, though future work may need to address long-tail performance through data augmentation or few-shot learning techniques.

6 Conclusion

This paper presented a baseline system for the AbjadNLP 2026 shared task on authorship identification for the Kurdish language, focusing on the Sorani dialect (Task 3). To the best of our knowledge, this is the first study addressing multiclass authorship identification for Kurdish within a shared-task setting. Our approach is based on fine-tuning the pretrained XLM-RoBERTa model and provides a strong baseline, achieving an accuracy of approximately 75% on the validation set. The reported results demonstrate the feasibility of applying multilingual transformer models to low-resource, Arabic-script languages such as Kurdish.

Several promising directions remain open for future work. These include exploring traditional machine learning and deep learning approaches tailored to stylistic features, investigating ensemble and hybrid models, and leveraging cross-lingual transfer from related high-resource languages such as Arabic to further improve authorship identification performance for Kurdish

References

Bakhtawar Abdalla, Rebwar Mala Nabi, Hassan Eshkiki, and Fabio Caraffini. 2025. Named entity recognition for the kurdish sorani language: Dataset creation and comparative analysis. *arXiv preprint arXiv:2511.22315*.

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.

Shakhawan Aghajan and Rebwar M. Nabi. 2025. Kusarcasm: Automated annotation of a sarcasm dataset using hybrid nlp techniques. *Data in Brief*, 63:112215.

Thamer Maseer Alharbi. 2025. Mission at arageneval shared task: Enhanced arabic authority classification. pages 14–17.

Soran Badawi. 2023. Kmd: A new kurdish multilabel emotional dataset for the kurdish sorani dialect. In *Proceedings of the 6th International Conference on Natural Language and Speech Processing (ICNLSP 2023)*, pages 308–315. Association for Computational Linguistics.

Soran Badawi. 2025. Hopedetect: a multicomponent deep learning framework for hope detection in kurdish language. *The Computer Journal*, 68:1743–1754.

Soran Badawi, Arefeh Kazemi, and Vali Rezaie. 2025a. Kurdisent: a corpus for kurdish sentiment analysis. *Language Resources and Evaluation*, 59:601–620.

Soran S. Badawi. 2024. Bridging the gap: Enhancing kurdish news classification with rfo-cnn hybrid model. *ARO-The Scientific Journal of Koya University*, 12:100–107.

Soran S Badawi, Ari M Saeed, Sara A Ahmed, and Diyari A Hassan. 2025b. Enhanced category-feature association measure: A robust approach for text classification through feature selection. *ARO-THE SCIENTIFIC JOURNAL OF KOYA UNIVERSITY*, 13(2):114–123.

Mena Hany. 2025. Nlp wizard at arageneval shared task: Embedding-based classification for ai detection and authorship attribution. pages 37–41.

Muhammad Helmy, Batool Balah, Ahmed Mohamed Sallam, and Ammar Sherif. 2025. Sebaweh at arageneval shared task: Berense-bert based ensembler for arabic authorship identification. pages 59–64.

Pshtiwan Jabar Karim and Karwan Osman Abdullah. 2025. A comprehensive study of machine learning and deep learning methods for sentiment analysis on kurdish sorani text. *Passer Journal of Basic and Applied Sciences*, 7:1118–1130.

Payman Sabr Rostam and Rebwar Mala Nabi. 2025. Bochun: Automatically annotated stance detection dataset for sorani kurdish language. *Data in Brief*, 61:111839.

Amr Sabaa and Mohamed Sabaa. 2025. Amr&mohamedsabaa at arageneval shared task: Arabic authorship identification using term frequency-inverse document frequency features with supervised machine learning. pages 32–36.

Eman Samir, Mahmoud Rady, Maria Bassem, Mariam Hossam, Mohamed Amin, Nisreen Hisham, Sara Gaballa, and Ayman Khalafallah. 2025. Athership at arageneval shared task: Identifying arabic authorship with a dual-model logit fusion. pages 54–58.

Rania Azad M. San Ahmed, Bilal Mohammed, Rawaz Mahmud, Lanya Zrar, and Shajwan Sdiq. 2021. Fake news detection in low-resourced languages" kurdish language" using machine learning algorithms. *Turkish Journal of Computer and Mathematics Education*, 12(6):4219–4225.

Rania Azad M. San Ahmed and Soran AB Saeed. 2025. Kurdabsa: Kurdish aspect-based sentiment analysis dataset curation using few-shot learning. *Data in Brief*, page 112012.

Chovyan H. Wahid and Rebwar M. Nabi. 2025. Adyan: automated annotating named entity recognition dataset for sorani kurdish language. *Data in Brief*, 62:111999.