Athership at AraGenEval Shared Task: Identifying Arabic Authorship with a Dual-Model Logit Fusion

Eman Samir*, **Mahmoud Rady***, **Maria Bassem***, **Mariam Hossam***, **Mohamed Amin***, **Nisreen Hisham***, **Sara Gaballa***
Applied Innovation Center (MCIT), Egypt
{e.samir, m.rady,m.bassem,mariam.hossam, m.amin,n.hisham,s.gaballa}@aic.gov.eg
Ayman Khalafallah
ayman.khalafallah@alexu.edu.eg

Abstract

Authorship identification in Arabic is a challenging task due to the language's morphological richness, orthographic variation, and stylistic diversity across genres and authors. In this paper, we present our submission to Subtask 2: Authorship Identification of the AraGenEval 2025 Shared Task at ArabicNLP, which aims to identify the author of a given Arabic paragraph among a set of 21 authors. This task is important for applications such as digital forensics, plagiarism detection, literary analysis, and AI-generated content verification, where reliably linking text to its author can provide critical insights. We employ transformer-based encoders and address the dataset's class imbalance by leveraging an ensemble of two capable Arabic language understanding models: AraBERT and AraELECTRA. Our approach combines the pre-softmax logits of both models before the final softmax layer, effectively capturing complementary strengths in their predictions. Using our proposed method, we achieved third place on the Subtask 2 leaderboard of the AraGenEval Shared Task (Abudalfa et al., 2025), with a Macro-F1 score of 0.85968 and accuracy of 0.89516 on the test split.

1 Introduction

This paper details the system we developed for the AraGenEval 2025 Shared Task on Arabic Authorship and AI-Generated Text Detection, hosted at the Arabic Natural Language Processing Conference (ArabicNLP 2025) (Abudalfa et al., 2025). Our work is submitted under Subtask 2: Authorship Identification, a multi-class classification challenge designed to attribute a given Arabic text to its correct author from a closed set of 21 distinguished writers. The importance of this task has grown substantially with the proliferation of digital content. Robust authorship identification systems have critical real-world applications in digital forensics for identifying anonymous authors, in cybersecurity for detecting coordinated disinformation campaigns, in academic integrity for uncovering plagiarism, and in digital humanities for attributing disputed or anonymous literary works. The task is centered exclusively on the **Arabic language**, with a dataset curated to include diverse genres such as literary, philosophical, and journalistic prose, ensuring that solutions must focus on deep stylistic features rather than superficial topical cues.

The challenge of authorship attribution in Arabic is particularly acute due to the language's intrinsic complexities. Arabic is characterized by its **rich and complex morphology**, where a single root can spawn a vast array of words, making traditional bag-of-words models less effective. Furthermore, the phenomenon of **diglossia**—the coexistence of Modern Standard Arabic (MSA) with numerous regional dialects—means that authors often possess a unique stylistic blend, which may not be immediately apparent. Finally, **orthographic variability** in the Arabic script, such as the multiple forms of the hamza and the optionality of diacritics (*tashkeel*), introduces surface-level noise that can obscure an author's true stylistic signature. These linguistic hurdles are compounded by difficulties inherent in the dataset itself, including a notable **class imbalance** across the authors and significant stylistic diversity. Together, these complexities demand robust models capable of identifying an author's unique textual fingerprint amidst considerable noise.

To address these challenges, we fine-tuned two state-of-the-art Arabic Transformer encoders: **AraBERT** (Antoun et al.), trained with Masked Language Modeling (MLM), and **AraELECTRA** (Antoun et al., 2021), trained with Replaced Token Detection (RTD). Their complementary pretraining objectives were expected to cap-

*These authors contributed equally to this work.

Proceedings of The Third Arabic Natural Language Processing Conference, pages 54–58

ture different facets of authorial style. Our best system is a logit-level ensemble that averages the models' raw prediction scores before the softmax, leveraging their strengths and reducing individual weaknesses. We also tested a sliding-window strategy with AraBERTv02 for handling inputs longer than 512 tokens.

Our ensemble-based system, achieved **3rd place** in the final competition rankings, demonstrating its effectiveness on this challenging task. The key contributions and findings of our work can be summarized as follows:

- We demonstrate the successful application of fine-tuned AraBERT and AraELECTRA models for Arabic authorship attribution, using minimal preprocessing to ensure the preservation of subtle stylistic markers.
- We show that a logit-level ensemble of AraBERT and AraELECTRA significantly outperforms either model individually on both the development and final test sets, confirming the value of model fusion.
- We provide a valuable negative result from our sliding-window experiments with AraBERTv02, which indicates that simple chunking and aggregation for documents longer than 512 tokens degrades performance, highlighting the critical importance of contiguous context for stylistic analysis.
- We present a qualitative analysis, including correctly classified examples from stylistically complex passages, to illustrate the system's practical capabilities.

2 Related Work

Authorship attribution has evolved from early stylometric methods based on lexical and statistical features (Stamatatos, 2009) to modern deep learning approaches. For Arabic, traditional machine learning methods using character n-grams and morphological features (Shaker, 2017; Haddad et al., 2019) have shown promise but require extensive feature engineering. Neural models such as RNNs and CNNs (Alshahrani and Alsuhaymi, 2020) reduce this need, and transformer-based encoders like AraBERT (Abdul-Mageed et al., 2021) and AraELECTRA (Antoun et al., 2021) now achieve state-of-the-art results in Arabic NLP. Ensemble methods remain underexplored for Arabic authorship tasks, with only limited work in social media contexts (Alshehri and Al-Khazraji, 2022), despite evidence from other languages (Jafari Akinabad and Mohammadpour, 2021) that model combination can improve robustness. Our work fills this gap by applying a logit-level ensemble of AraBERT and AraELECTRA for literary and philosophical genres.

3 Dataset

The dataset was curated by the task organizers from 10 publicly available books for 21 authors. Books were segmented into semantically coherent paragraphs, yielding substantial variation in length and style. Table 1 summarizes the distribution of samples across train, validation, and test splits.

Author	Train	Val	Test
Ahmed Amin	2892	246	594
Ameen Rihani	1557	142	624
Hassan Hanafi	3735	548	1002
...	...	...	...
William Shakespeare	1236	238	358

Table 1: Example excerpt of dataset statistics; full table provided by organizers.

Paragraph lengths range from short excerpts of under 50 tokens to long passages exceeding the 512-token limit of standard Transformer models. The dataset is also **imbalanced**, with author sample counts ranging from a few hundred to several thousand, introducing a challenge for models to maintain performance on minority classes.

4 Methodology

4.1 Base Models: AraBERT and AraELECTRA

AraBERT is a 12-layer bidirectional Transformer encoder based on BERT (Devlin et al., 2019), pretrained on large-scale Arabic corpora (news, Wikipedia, social media) using the Masked Language Modeling (MLM) objective. This bidirectional training captures deep contextual relationships between words and morphemes, beneficial for Arabic's rich morphology. For our task, we add a linear classification layer on the final hidden state of the [CLS] token.

AraELECTRA follows the ELECTRA framework (Clark et al., 2020), replacing MLM with a Replaced Token Detection (RTD) objective, where

the model discriminates between original and substituted tokens. This more sample-efficient training yields rich token-level representations. Architecturally, it is also a 12-layer Transformer encoder, with the same classification head as AraBERT.

We fine-tune both models for 4 epochs with a maximum sequence length of 512 tokens, truncating longer texts. This identical setup enables direct comparison and facilitates their combination in our logit-level ensemble.

4.2 Logit-Level Ensemble

Each model outputs logits $\ell^{(1)}, \ell^{(2)} \in R^{11}$. We combine them as:

$$\ell_{\text{ens}} = \ell^{(1)} + \ell^{(2)}, \quad p = \text{softmax}(\ell_{\text{ens}})$$

This preserves raw decision margins before applying the softmax.

4.3 Sliding-Window Experiment

We fine-tuned the BERT Large AraBERTv02 model (aubmindlab/bert-large-AraBERTv02) for authorship identification using a sliding-window approach to handle long paragraphs without losing context. Input texts were split into fixed-length sequences of 512 tokens (including special tokens) with a stride of 128 tokens, ensuring overlap between adjacent segments so that stylistic cues spanning boundaries were preserved.

The dataset was loaded from Excel files with author names label-encoded. To address class imbalance, balanced class weights were computed and passed to a custom Trainer subclass. We applied label smoothing with a factor of 0.1 to improve generalization.

At inference, document-level voting was implemented by aggregating chunk predictions to produce the final author label.

4.4 Baselines

- TF-IDF + FCN: Character and word n-gram features via TF-IDF, fed into a 2-layer fully connected network.

- Contrastive (Qarib) + k-NN: Contrastive learning on Qarib (Abdelali et al., 2021) encoder embeddings to bring same-author texts closer in vector space, followed by k-nearest neighbors classification.

4.5 Negative Experiment: Simple Chunking

We attempted to split long texts (>512 tokens) into smaller chunks (512 and remainder), assigning the same label to all chunks. This degraded accuracy, likely because shorter fragments sometimes lack sufficient stylistic cues.

5 Results

Table 2 summarizes the performance of our models on the development set. Among the individual models, **AraBERT** achieved the highest development accuracy (0.90) and a Macro-F1 score of 0.84, slightly outperforming AraELECTRA (0.88 accuracy, 0.83 Macro-F1). Our logit-level **ensemble** of AraBERT and AraELECTRA produced the best overall results on the development set, with an accuracy of 0.92 and a Macro-F1 score of 0.86, confirming the benefit of combining the two architectures.

We also evaluated several alternative approaches. A **sliding-window** inference strategy applied to AraBERTv02 slightly improved the Macro-F1 score over the single-model baselines (0.85) but did not surpass the ensemble. Traditional **TF-IDF** features followed by a fully connected network (FCN) performed considerably worse (approximately 0.75 accuracy, 0.70 Macro-F1), highlighting the limitations of shallow lexical representations for this task. A **contrastive learning** approach achieved moderate performance (0.84 accuracy, 0.79 Macro-F1), suggesting that more specialized contrastive objectives might be needed for stylistic analysis.

Our final submission to the AraGenEval 2025 Subtask 2 leaderboard achieved a Macro-F1 score of 0.85968 and an accuracy of 0.89516 on the held-out test set, placing third overall in the competition. These results demonstrate the effectiveness of our ensemble strategy in capturing complementary stylistic cues from the two pretrained models.

To illustrate the system's ability to capture nuanced stylistic patterns, we present two correctly classified examples from the test set:

Example 1 — Philosophical Prose:
Input excerpt: وهنا تصبح الطبيعة في حاجة إلى مبرر وهكذا يقدم كانت فرضًا تفسيريًا محضًا لا يغير من محتوى المعرفة.

Model	Dev Acc.	Dev Macro F1
AraBERT	0.90	0.84
AraELECTRA	0.88	0.83
Ensemble	0.92	0.86
Sliding Window	0.90	0.85
TF-IDF + FCN	∼0.75	0.70
Contrastive	0.84	0.79

Table 2: Model performance on the development set.

Predicted author: فؤاد زكريا *(correct)*

Example 2 — Literary Prose:

Input excerpt: ثمن الكتابة... لا أجيد كتابة المقدمات، يمكن أن أكتب قصةً من ألف صفحة... يدق بالمطرقة على جواز سفرها فتدخل.

Predicted author: نوال السعداوي *(correct)*

6 Discussion

The experimental results indicate that the ensemble of AraBERT and AraELECTRA consistently outperformed either model individually on both the development and test sets. We attribute this improvement to the complementary nature of the models' pretraining objectives: AraBERT's masked language modeling encourages deeper bidirectional context modeling, while AraELECTRA's replaced token detection promotes fine-grained token-level discrimination. By combining their pre-softmax logits, the ensemble is able to integrate these distinct strengths, leading to more robust stylistic representation and classification.

The limited gains observed from the sliding-window approach suggest that splitting long texts into chunks may disrupt important discourse-level cues, which are often essential for capturing an author's style. Similarly, the relatively low performance of the TF-IDF + FCN baseline confirms that surface lexical features alone are insufficient for distinguishing between highly skilled Arabic authors with overlapping vocabularies. The moderate results of the contrastive learning approach point to the need for more task-specific contrastive objectives that explicitly model stylistic similarity and difference.

Overall, the findings highlight the value of leveraging multiple pretrained encoders with different inductive biases, while also underscoring the importance of preserving global context in Arabic authorship attribution tasks.

7 Conclusion and Future Work

In this paper, we presented a logit-level ensemble of AraBERT and AraELECTRA for Arabic authorship attribution, developed for the AraGenEval 2025 Shared Task. Our approach leveraged the complementary strengths of two transformer-based encoders with different pretraining objectives, resulting in robust performance across literary, philosophical, and journalistic genres. The system achieved third place on the competition leaderboard, with a Macro-F1 score of 0.85968 and an accuracy of 0.89516 on the held-out test set. The results demonstrate that combining pretrained models is an effective strategy for addressing the linguistic and stylistic challenges of Arabic authorship identification.

For future work, we plan to extend our ensemble in two directions. First, we will explore *weighted* logit-level fusion, where the contribution of each model is learned or tuned based on validation performance rather than averaged equally. Second, we aim to increase the number of diverse models in the ensemble, incorporating additional pretrained Arabic encoders and possibly multilingual transformers. We expect that both strategies will further enhance performance by capturing a wider range of stylistic and contextual features, thereby improving the system's robustness and generalization.

Acknowledgments

We would like to thank the organizers of the AraGenEval 2025 Shared Task for providing the dataset and evaluation platform, and for their efforts in fostering research on Arabic NLP. Finally, we are grateful to our colleagues and peers for their valuable feedback during the development of this work.

References

Ahmed Abdelali, Sabit Hassan, Hamdy Mubarak, Kareem Darwish, and Younes Samih. 2021. Pre-training bert on arabic tweets: Practical considerations.

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7088–7105, Online. 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 (ArabicNLP 2025)*. Association for Computational Linguistics.

Nourah Alshahrani and Hadeel Alsuhaymi. 2020. Deep learning for arabic authorship attribution. In *2020 6th International Conference on Information Management (ICIM)*, pages 8–14. IEEE.

Nourah Alshehri and Mohammed Al-Khazraji. 2022. Ensemble methods for arabic author profiling on social media data. *Future Internet*, 14(2):51.

Wissam Antoun, Fady Baly, and Hazem Hajj. Arabert: Transformer-based model for arabic language understanding. In *LREC 2020 Workshop Language Resources and Evaluation Conference 11–16 May 2020*, page 9.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2021. AraELECTRA: Pre-training text discriminators for Arabic language understanding. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 191–195, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Ramy Haddad, Wael Obeid, and Hani Jundi. 2019. Morphological features for arabic authorship attribution. In *International Conference on Information and Communication Technologies for Development*, pages 615–624. Springer.

Zahra Jafari Akinabad and Masoud Mohammadpour. 2021. Ensemble learning methods in authorship attribution. In *2021 7th International Conference on Web Research (ICWR)*, pages 1–6. IEEE.

Zaid Shaker. 2017. Arabic authorship identification using n-gram and support vector machine. In *International Conference on Advanced Machine Learning Technologies and Applications*, pages 507–516. Springer.

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

A Appendix

A.1 Hyperparameter Settings

Table 3 lists the main hyperparameters used for fine-tuning AraBERT and AraELECTRA in our experiments.

Hyperparameter	**Value**
Optimizer	AdamW
Learning rate	2×10^{-5}
Batch size	16
Weight decay	0.01
Epochs	4
Max sequence length	512

Table 3: Hyperparameters for fine-tuning.

A.2 Hardware and Runtime

All experiments were run on a single NVIDIA P100 GPU with 16GB of memory. Fine-tuning each model for 4 epochs required approximately 3 hours.