HCMUS_PrisonDilemma at AbjadAuthorID Shared Task: Less is More with Base Models

Huynh Trung Kiet[1,3*], Dao Sy Duy Minh[1,3*], Tran Chi Nguyen[1,3],
Nguyen Lam Phu Quy[1,3], Pham Phu Hoa[1,3], and Truong Bao Tran[2,3]

[1]Faculty of Information Technology, University of Science, Ho Chi Minh, Vietnam
[2]University of Economics and Law, Ho Chi Minh, Vietnam
[3]Vietnam National University, Ho Chi Minh City, Vietnam

{23122039, 23122041, 23122044, 23122048, 23122030}@student.hcmus.edu.vn
trantb234102e@st.uel.edu.vn

Abstract

We present our approach to the AbjadNLP 2026 Arabic Authorship Identification shared task, achieving 4th place. Our key finding is that AraBERT-base (110M) outperforms AraBERT-large (340M) on the test set with macro F1 of 0.8449 versus 0.8096, despite lower validation scores. We handle long passages via sliding window chunking with mean pooling, and use a two-stage classification head with dual dropout for regularization. Per-class analysis reveals that translated works achieve perfect F1 while classical poets remain challenging due to shared formal structures. Our results challenge the "scale is all you need" assumption for stylometric tasks.

1 Introduction

"The style is the man himself." This famous aphorism by Georges-Louis Leclerc captures a timeless truth that has fascinated scholars for centuries: every writer leaves an indelible stylistic fingerprint in their prose (Stamatatos, 2009). From the rhythm of sentences to the choice of function words, from syntactic preferences to punctuation habits, these unconscious patterns constitute a literary DNA that persists across topics and genres (Koppel et al., 2009).

The ABJADAUTHIDEN shared task (Abudalfa et al., 2025, 2026) challenges participants to distinguish between 21 prominent Arabic authors spanning philosophers, novelists, poets, and translators. The task presents unique challenges: significant class imbalance with a 9:1 ratio between the largest and smallest classes, long literary passages exceeding standard transformer context windows (Devlin et al., 2019), and the inherent difficulty of distinguishing authors who share similar genres and historical periods.

Our team, **HCMUS_PrisonDilemma**, initially hypothesized that larger pre-trained language models would capture more nuanced stylistic patterns, following the conventional wisdom established by scaling studies in modern NLP (Vaswani et al., 2017). However, our investigation revealed a counter-intuitive finding: the base variant of AraBERT (Antoun et al., 2020) consistently outperformed its larger counterpart on the held-out test set, despite achieving lower validation scores. This pattern suggests differential overfitting (Srivastava et al., 2014), where the larger model's capacity led to memorization of validation-specific patterns rather than acquisition of generalizable stylistic features. We note this observation emerges from a single model pair comparison, and broader scaling conclusions would require more extensive experimentation across multiple architectures and random seeds.

Our contributions are threefold. First, we demonstrate empirically that AraBERT-base outperforms AraBERT-large for this task, challenging assumptions about model scaling for stylometry. Second, we provide detailed per-class analysis revealing that genre and stylistic distinctiveness strongly correlate with classification performance. Third, we present a resource-efficient training strategy optimized for standard GPU environments that achieves competitive results without requiring expensive computational infrastructure.

2 Related Work

Authorship attribution traces its origins to Mosteller and Wallace's analysis of the Federalist Papers, where function-word frequencies proved effective for author disambiguation (Stamatatos, 2009). Subsequent work expanded the feature space to include character n-grams and the "writeprints" framework (Koppel et al., 2009; Abbasi and Chen, 2008). The deep learning revolution enabled end-to-end learning of stylistic features, with pre-trained language models now dominating benchmarks (Devlin et al., 2019).

For Arabic NLP, AraBERT (Antoun et al., 2020)

Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script (AbjadNLP 2026), pages 448–452

pioneered Arabic BERT variants through pre-training on news and web corpora. CAMeLBERT (Inoue et al., 2021) analyzed how pre-training domains affect downstream performance, while MARBERT and ARBERT (Abdul-Mageed et al., 2021) expanded the ecosystem with social media and formal text variants. A key challenge is handling documents exceeding BERT's 512-token limit; we adopt sliding window chunking for computational efficiency.

3 Task and Dataset

The shared task requires multiclass classification into 21 author categories (Abudalfa et al., 2026). The corpus comprises 47,692 passages from literary works: 35,122 training, 4,157 validation, and 8,413 test samples, with passages extending up to approximately 1,900 tokens per sample. Each book was segmented into semantically coherent paragraphs by the task organizers.

A distinctive feature of this dataset is the organizers' decision to employ GPT-4o mini to rephrase selected paragraphs into a standardized formal style (Abudalfa et al., 2025). This design choice aims to normalize surface-level variation and focus the attribution challenge on deeper stylistic patterns rather than incidental formatting differences. However, such paraphrasing may inadvertently alter certain authorial fingerprints, representing an inherent characteristic of the official dataset that all participating teams navigated.

The 21 authors span remarkable diversity across genre, era, and stylistic register. Philosophers like Hassan Hanafi and Fouad Zakaria employ dense terminological prose laden with specialized vocabulary. Novelists including Nobel laureate Naguib Mahfouz craft distinctive narrative voices through dialogue and storytelling rhythms. Poets such as Ahmed Shawqi maintain classical formal structures with recognizable metrical patterns. Translators like Robert Barr introduce unique lexical patterns from Western source texts, including loan words and sentence structures influenced by the original language. We used the official dataset without additional filtering of author names or book titles. Such metadata, if present, could constitute information leakage, particularly for translators whose works may retain source attribution. Figure 1 illustrates the training distribution, showing pronounced class imbalance.

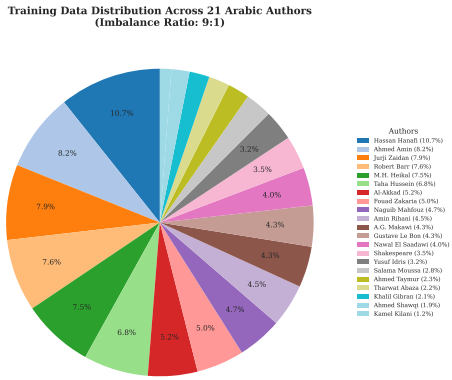

Figure 1: Training data distribution across 21 authors showing 9:1 class imbalance ratio.

4 System Description

Our system is built upon the HuggingFace Transformers library (Wolf et al., 2019) and optimized for the Kaggle T4 GPU environment with 16GB of memory. The complete pipeline consists of three main components: a sliding window mechanism for handling long documents, a pre-trained Arabic encoder backbone, and a custom two-stage classification head with regularization. Figure 2 illustrates the end-to-end architecture.

Long-Context Strategy: Literary passages in this dataset frequently exceed AraBERT's maximum context window of 512 tokens (Devlin et al., 2019), with some samples approaching 1,900 tokens. To address this challenge, we implement a sliding window approach with window size $L = 512$ tokens and stride $S = 256$ tokens, producing overlapping chunks that preserve local context at boundaries. Each chunk is independently encoded by the transformer backbone, and we extract the [CLS] token embedding from each chunk. These chunk-level representations are then aggregated via mean pooling to form a single 768-dimensional document representation. This strategy ensures complete coverage of long documents while maintaining computational efficiency on our resource-constrained environment.

Encoder Backbone: We employ `aubmindlab/bert-base-arabertv02` (Antoun et al., 2020), a 12-layer transformer with 768 hidden units and 110M parameters, pre-trained on approximately 70GB of Arabic news and web text. This model was selected based on preliminary experiments where we also evaluated MARBERT

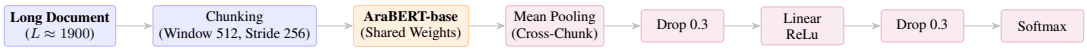

Figure 2: System architecture with sliding window chunking and mean pooling aggregation.

(Abdul-Mageed et al., 2021) and CAMeLBERT (Inoue et al., 2021). AraBERT-base achieved marginally better validation performance on this literary text domain, though we acknowledge these comparisons were conducted under time constraints without exhaustive hyperparameter tuning for each model.

Classification Head: Rather than a single linear layer, we implemented a two-stage classification head designed to capture non-linear relationships in the stylistic embedding space (Srivastava et al., 2014). The architecture proceeds as: Dropout(0.3) $\rightarrow$ Linear(768$\rightarrow$512) $\rightarrow$ ReLU $\rightarrow$ Dropout(0.3) $\rightarrow$ Linear(512$\rightarrow$21) $\rightarrow$ Softmax. This dual-dropout design proved important for regularization, preventing the model from overfitting to superficial patterns in the training data.

Training Configuration: We employed the AdamW optimizer (Loshchilov and Hutter, 2019) with learning rate 2×10^{-5} and weight decay 0.01, using a linear learning rate scheduler with 10% warmup steps. To achieve effective batch sizes without memory overflow on our T4 GPU, we utilized gradient accumulation with 2 accumulation steps and a base batch size of 64, yielding an effective batch size of 128 samples. Gradient clipping with maximum norm 1.0 prevented exploding gradients during training. Models were trained for 10 epochs with early stopping patience of 3 epochs based on validation macro F1.

5 Experiments

5.1 Official Leaderboard Results

Our submission achieved 4$^{\text{th}}$ place on the official leaderboard among all participating teams. Table 1 presents top team performances as recorded at the shared task deadline.

Rank	Team	F1-Score	Accuracy
1	zaghloul2012	0.9185	0.9571
2	grkurdi	0.8897	0.9244
3	33_tree	0.8696	0.9050
4	**HCMUS_PrisonDilemma**	**0.8449**	**0.8767**
5	mayar_boghdady	0.8400	0.8804
6	shahadsuh	0.8077	0.8776
7	Ali Al-Laith	0.7918	0.8479
8	hurryte	0.7901	0.8300

Table 1: Official leaderboard results on the test set.

5.2 Model Scale Comparison

Table 2 presents our internal experiments comparing different model configurations. The base model (110M parameters) outperforms the large model (340M parameters) on the test set by 3.5 percentage points in macro F1, while this pattern is reversed on validation—a signature characteristic of overfitting (Srivastava et al., 2014).

We interpret this finding through the lens of stylometric signal complexity: authorship attribution relies on relatively consistent patterns of word choice, syntactic preference, and rhythmic tendency that may be adequately captured by smaller models (Stamatatos, 2009), while excess capacity enables memorization of spurious correlations. However, this interpretation derives from a single model pair without variance estimates, and broader claims would require more experimentation.

Model	Val F1	Test F1
AraBERT-base (5 ep.)	0.7119	0.7349
AraBERT-large (10 ep.)	**0.8405**	0.8096
AraBERT-base (10 ep.)	0.8287	**0.8449**

Table 2: Base model generalizes better despite lower validation performance.

6 Per-Class Analysis

Figure 3 reveals substantial variation in classification difficulty across author categories. The top-performing classes share distinctive stylistic traits that set them apart from the remainder of the corpus. Robert Barr achieves perfect F1 (1.00), reflecting unique lexical signatures of translated works. This exceptional performance warrants scrutiny: perfect identification may partially reflect artifacts such as

retained translator attributions or foreign named entities rather than pure stylistic recognition (Koppel et al., 2009).

Hassan Hanafi achieves F1 of 0.98 with his dense philosophical prose characterized by specialized terminology. Conversely, classical poets like Ahmed Shawqi prove challenging (F1=0.38) due to shared formal structures and metrical patterns that blur distinctions for semantic models trained primarily on prose (Stamatatos, 2009).

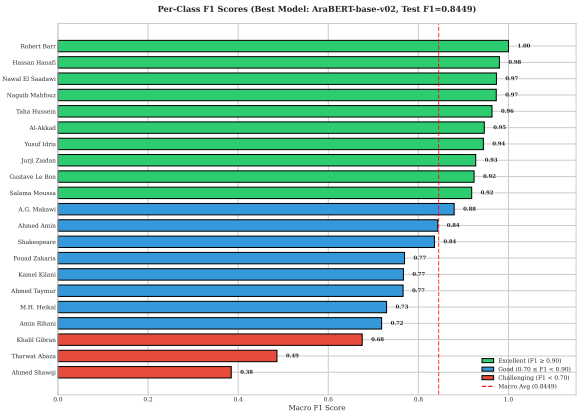

Figure 3: Per-class F1 scores: green ($\geq$0.90), blue (0.70–0.89), red ($<$0.70).

7 Training Dynamics

Figure 4 illustrates training dynamics of our base model configuration. Validation F1 jumps to 0.48 in epoch 1 as the model learns broad genre distinctions, reaches 0.70 by epoch 3, and continues improving to 0.83 by epoch 10 as fine-grained author distinctions are acquired.

We hypothesize that the larger model's inferior test performance stems from its enhanced capacity to memorize validation-specific patterns during the extended training regime (Srivastava et al., 2014). The base model's constrained capacity forces more generalizable features, though this interpretation remains speculative without controlled experiments.

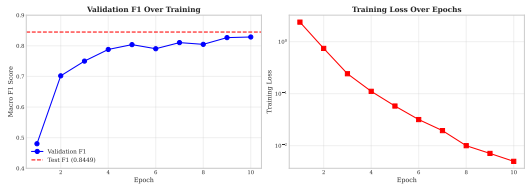

Figure 4: Training loss and validation F1 over 10 epochs.

8 Conclusion

We presented HCMUS_PrisonDilemma's approach for the ABJADAUTHIDEN shared task, demonstrating that AraBERT-base outperforms its larger counterpart on this benchmark. Our 4th place finish validates well-tuned base models for stylometric tasks. Future work should address ablation studies, multi-seed experiments, non-transformer baselines (Abbasi and Chen, 2008), potential label leakage investigation, and class-imbalance mitigation strategies.

Limitations

Our study has several limitations. All metrics derive from single training runs without multiple seeds, meaning observed differences may partially reflect initialization variance. We did not conduct systematic ablations to isolate contributions of the two-stage head or dropout rates. We did not explore class-imbalance mitigation (class weighting, focal loss). Potential label leakage through author names or translator attributions was not investigated. Our comparison focused exclusively on AraBERT variants without comprehensive evaluation of other Arabic PLMs (Inoue et al., 2021; Abdul-Mageed et al., 2021) or non-transformer baselines (Koppel et al., 2009). Finally, generalization beyond literary texts remains untested.

Ethics Statement

All texts derive from publicly accessible literary works. We acknowledge dual-use concerns: techniques for literary analysis could potentially be repurposed for surveillance or de-anonymization (Abbasi and Chen, 2008). We encourage appropriate access controls when deploying authorship attribution systems.

Acknowledgements

We would like to express our sincere gratitude to the organizers of the AbjadNLP 2026 workshop and the ABJADAUTHIDEN shared task for providing this valuable opportunity to advance research in Arabic authorship identification. We thank Dr. Shadi Abudalfa and the organizing committee for their efforts in curating the dataset and managing the competition. We also acknowledge Kaggle for providing the computational resources that made our experiments possible.

References

Ahmed Abbasi and Hsinchun Chen. 2008. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. In *ACM Transactions on Information Systems*, volume 26, pages 1–29.

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*, pages 7088–7105.

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.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools*, pages 9–15.

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*, pages 4171–4186.

Go Inoue, Bashar Alhafni, Ramy Baly, and Nizar Habash. 2021. The interplay of variant, size, and task type in arabic pre-trained language models. *arXiv preprint arXiv:2103.06678*.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2009. Computational methods in authorship attribution. In *Journal of the American Society for Information Science and Technology*, volume 60, pages 9–26.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.