QalamID at AbjadAuthorID Shared Task: Morphology Matters, A Hybrid Ensemble for Arabic Authorship Attribution

Youssef Zaghloul
Università Cattolica del Sacro Cuore / Milan, Italy
Universidad de Zaragoza / Zaragoza, Spain
youssef.zaghlouleweis01@icatt.it

Abstract

Arabic authorship attribution presents unique challenges due to the language's rich derivational morphology, which often fragments word-level frequencies. In this paper, we describe our winning submission to the AbjadAuthorID Shared Task. We propose a hybrid ensemble system that fuses the morphological precision of character n-gram LinearSVCs with the semantic understanding of fine-tuned Transformers (AraBERT and XLM-RoBERTa). Contrary to current trends in NLP, we demonstrate that traditional character n-grams (0.92 F1) significantly outperform deep learning baselines (AraBERT 0.87 F1) for this task, suggesting that authorial signature in Arabic is encoded more densely in morphological patterns than in semantic content. Our final system employs a novel "Precision Scalpel" post-hoc calibration technique and selective pseudo-labeling to address class imbalance and genre confounds. The system achieved the **1st place ranking** with a macro F1-score of **0.932** and accuracy of **0.963** on the test set.

1 Introduction

Authorship Attribution (AA), a task of identifying the author of a text from a closed set of candidates, remains a critical problem in forensic linguistics and digital humanities. While recent advances in Natural Language Processing (NLP) have been dominated by large pre-trained Transformers (Devlin et al., 2019; Antoun et al., 2020), applying these models to morphologically rich languages like Arabic presents distinct challenges. Arabic's root-and-pattern morphology generates over 10,000 verb forms per root, leading to high vocabulary sparsity that complicates standard stylometric approaches.

In this paper, we present the system developed by **QalamID** for the AbjadAuthorID Shared Task (Abudalfa et al., 2025, 2026). The task involves identifying 21 authors from a dataset of classical and modern Arabic literature, characterized by significant class imbalance (1:9 ratio) and "translationese" artifacts in non-native texts.

Our primary contribution is a rigorous demonstration that for Arabic AA, **morphology outperforms semantics**. Through extensive ablation studies, we found that a lightweight LinearSVC trained on character n-grams achieved a macro F1-score of 0.92, significantly outperforming a fine-tuned AraBERT v2 (0.87 F1). We hypothesize that while Transformers capture *what* an author says (topic), character n-grams implicitly capture *how* they say it (derivational suffixes, clitic attachments, and conjugation habits), signals that are more robust to topic shifts and genre changes.

To leverage the best of both paradigms, we developed a hybrid ensemble architecture. Our system combines the high-precision morphological signal of traditional models with the semantic, translation-aware signal of Transformers (AraBERT and XLM-RoBERTa). To address specific error modes, such as the confusion between the dramatist Tharwat Abaza and translated Shakespeare, we introduce two engineering innovations:

1. **The Consultant Strategy:** A constrained ensemble weighting method that retains weak-but-orthogonal models (such as syntactic POS n-grams) to resolve tie-breaking scenarios.

2. **Precision Scalpel:** A post-hoc calibration technique that scales class probabilities based on confusion matrix analysis, improving recall for minority classes by up to 23 F1 points without retraining.

Our system achieved the top rank on the shared task leaderboard. We provide a detailed analysis of why simple features remain competitive in the deep learning era and offer insights into the linguistic markers of Arabic literary style.

Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script (AbjadNLP 2026), pages 476–482

2 Background

2.1 Task Setup and Dataset

The AbjadAuthorID Shared Task focuses on authorship attribution for Arabic text. Formally, given a document d, the goal is to classify it into one of $C = 21$ author classes $A = \{a_1, ..., a_{21}\}$. The dataset comprises literary texts spanning classical and modern Arabic, as well as translated works, presenting significant linguistic diversity (Abudalfa et al., 2026).

The provided dataset consists of training, validation, and test splits. A critical preprocessing step in our pipeline was the identification and removal of duplicate samples. We detected that 6.5% of the original training data consisted of duplicates, which we removed to prevent model memorization. The final statistics of the dataset used in our experiments are summarized in Table 1.

Split	Samples
Original Training	35,122
After Deduplication	**32,748**
Validation	4,157
Test (Unlabeled)	8,413

Table 1: Dataset statistics after preprocessing.

Challenges The dataset poses three primary challenges:

1. **Class Imbalance:** The distribution is highly skewed (ratio 1:9), with the most frequent author (Hassan Hanafi) having 548 validation samples, while the least frequent (Kamel Kilani) has only 25.

2. **Text Length:** Documents average 316 words but reach up to 1,850 words (≈2,400 tokens), exceeding the standard 512-token limit of Transformer models.

3. **Genre Confounding:** The corpus contains a mix of genres (prose, poetry, drama). As discussed in Section 5, we observed that dramatic style (e.g., plays by Tharwat Abaza) is frequently confused with translated dramatic works (e.g., Shakespeare), creating a "genre confound" where models struggle to disentangle authorial style from genre conventions.

2.2 Related Work

Authorship attribution has historically relied on stylometric features such as function words and Burrows' Delta (Burrows, 2002). However, applying standard Delta to Arabic is complicated by the language's rich morphology, where clitics and affixes fragment word frequencies. While explicit morphological segmentation can mitigate this, our preliminary experiments showed that standard stylometric features (0.69 F1) underperformed compared to n-gram baselines.

In recent years, deep learning approaches, particularly pre-trained Transformers like BERT (Devlin et al., 2019) and AraBERT (Antoun et al., 2020), have achieved state-of-the-art results in text classification. While powerful, these models primarily encode semantic information. Our work challenges the assumption that semantic embeddings are superior for authorship tasks. We demonstrate that for Arabic, character-level n-grams (Stamatatos, 2009), which implicitly capture root-and-pattern morphology, provide a stronger signal (0.92 F1) than AraBERT (0.87 F1), motivating our proposed hybrid architecture.

3 System Overview

Our winning system is a hybrid ensemble that integrates five distinct models across two paradigms: traditional feature-based classifiers (SVMs) and fine-tuned Transformers. The architecture is designed to capture authorship signals at three levels: morphological (character n-grams), lexical (word n-grams), and syntactic (POS n-grams).

3.1 Text Preprocessing Pipeline

We adopted a conservative preprocessing strategy to preserve stylistic signals. Our pipeline applies orthographic normalization (unifying Alif forms and Yaa/Alif Maqsura) and removes Tatweel (elongation). Crucially, we strip diacritics (tashkeel) only *after* any morphological analysis to prevent feature fragmentation. Unlike standard NLP pipelines, we explicitly retain stopwords, as function words serve as the "neural pathways" of authorship style (e.g., preference for *fi hin anna* "whereas" vs. *baynama* "while"). Finally, we removed 4,676 duplicate samples identified in the training and validation data to prevent memorization.

3.2 Branch 1: Traditional Feature Extraction

We developed a three-branch feature system fed into Calibrated LinearSVCs ($C = 1.0$, balanced class weights).

1. Morphological Branch (Character N-Grams): This is our primary performer (0.92 F1). We extract character n-grams ($n \in [2, 5]$) within word boundaries using a TF-IDF vectorizer (50k features, sublinear scaling). This approach implicitly captures Arabic's root-and-pattern morphology—detecting author-specific preferences in derivational suffixes and clitic attachments—without requiring error-prone explicit morphological segmentation.

2. Lexical Branch (Word N-Grams): We extract word n-grams ($n \in [1, 3]$) to capture vocabulary choices and collocations (30k features). While less robust to morphological variation than the character branch, it captures distinctive domain vocabulary essential for distinguishing philosophers.

3. Syntactic Branch (POS N-Grams): To capture sentence structure independent of topic, we implement a "Syntactic X-Ray" strategy. We use Stanza (Qi et al., 2020) to generate Universal POS tags for the first 1,500 characters of each text, then extract POS n-grams ($n \in [3, 5]$). While this model's standalone performance is lower (0.67 F1), it acts as a "Consultant" in our ensemble, resolving confusion between authors who share vocabulary but differ in sentence complexity (e.g., the Philosopher Cluster).

3.3 Branch 2: Deep Learning Models

We fine-tuned two Transformer models: **AraBERT v2** (for Arabic semantic nuance) and **XLM-RoBERTa** (for detecting translation artifacts). To handle document lengths exceeding 512 tokens, we implemented a hierarchical sliding window strategy:

1. **Chunking:** Documents are split into 512-token chunks with a 256-token stride (50% overlap).

2. **Hierarchical Attention:** Instead of mean-pooling, we learn an attention mechanism over chunk representations (h_{CLS}) to compute a weighted document embedding D:

$$\alpha_i = \text{softmax}(w^T \tanh(W h_{CLS}^{(i)})) \quad (1)$$

$$D = \sum \alpha_i h_{CLS}^{(i)} \quad (2)$$

This allows the model to prioritize stylistically dense sections (e.g., conclusions) over generic introductions.

3.4 Ensemble and Calibration

Constrained Voting (The Consultant Strategy): We aggregate the probability distributions of the five models using a weighted soft voting scheme. Weights were optimized via random search with a domain constraint: the Syntactic branch weight was capped at 15%. This ensures the ensemble benefits from syntactic diversity without being overwhelmed by the branch's lower individual accuracy.

Precision Scalpel (Post-Hoc Calibration): To address the 1:9 class imbalance and specific confusion patterns (e.g., Abaza vs. Shawqi), we introduce the *Precision Scalpel*. This technique adjusts the ensemble's output probabilities $P(y|x)$ for specific classes c using a scalar multiplier λ_c:

$$P_{calibrated}(y = c|x) = P(y = c|x) \cdot \lambda_c \quad (3)$$

Multipliers are determined by analyzing the validation confusion matrix: classes with high precision but low recall (like Tharwat Abaza) receive a boost ($\lambda > 1.0$), while "gravity well" classes that attract false positives (like Ahmed Shawqi) are penalized ($\lambda < 1.0$). This simple calibration improved recall for challenging authors by up to 23 points.

4 Experimental Setup

Data Splits We utilized the official shared task splits: 30,446 training samples (after deduplication), 4,157 validation samples, and 8,413 unlabeled test samples.

Semi-Supervised Learning (Pseudo-Labeling) To align the training distribution with the test set, we employed a selective pseudo-labeling strategy. We generated predictions on the unlabeled test set using our initial ensemble. High-confidence predictions (probability ≥ 0.95) were treated as ground truth and added to the training set (adding 3,360 samples). Crucially, to maintain computational efficiency, we only retrained the lightweight LinearSVC branches on this augmented data, leaving the expensive Transformer models static.

Implementation Details Traditional models were implemented using `scikit-learn` with Platt scaling for probability calibration. Transformers were fine-tuned using HuggingFace `transformers` for 3 epochs (AraBERT) and 4 epochs (XLM-R) with a batch size of 8 and a learning rate of 2e-5. All experiments were conducted on a single P100 GPU. Hyperparameters are detailed in Appendix A.

5 Results

5.1 Quantitative Performance

Our final system achieved a Macro F1-score of 0.948 on the validation dataset and maintained a Macro F1-score of **0.932** and accuracy of **0.963** on the test set, securing the **1st place ranking**. Table 2 presents the performance of individual components and the final ensemble. For a comprehensive breakdown of all single-model baselines, detailed ensemble configurations, and full per-class performance metrics, refer to Appendix B.

Model Configuration	Val F1	Time
Baselines		
Syntactic Branch (POS N-Grams)	0.67	1m
AraBERT v2 (Fine-tuned)	0.87	2h
Word N-Grams (Lexical)	0.89	2m
Char N-Grams (Morphological)	**0.92**	**3m**
Ensembles		
Ensemble (Uncalibrated)	0.936	-
+ Pseudo-Labeling	0.941	+5m
+ Precision Scalpel (Final)	**0.948**	-

Table 2: Ablation study showing Validation F1 and approximate training time. Character N-Grams outperform AraBERT while being $40\times$ faster to train.

5.2 Ablation Analysis

Morphology vs. Semantics: A key finding of this work is that the Character N-Gram model (0.92 F1) significantly outperforms AraBERT (0.87 F1). This confirms our hypothesis that Arabic authorship is encoded more densely in morphological patterns (captured by character n-grams) than in semantic topics (captured by BERT).

Computational Efficiency: The hybrid approach offers a favorable trade-off between accuracy and cost. As shown in Table 2, the SVM branches can be retrained in minutes (e.g., for pseudo-labeling iterations), whereas the BERT models require hours. By restricting pseudo-labeling updates to the SVMs, we achieved a +1.0% improvement on the Test F1 with negligible added compute. Full details on feature counts and training resources for all models are listed in Table 6 (Appendix B).

5.3 Error Analysis and Calibration

The "Precision Scalpel" proved critical for challenging classes. Detailed qualitative analysis of these error patterns is provided in Table 9 (Appendix C), and visualizations of the baseline performance prior to calibration are available in Appendix D.

- **Tharwat Abaza:** Initially, models confused Abaza's plays with translated Shakespeare due to shared dramatic conventions (dialogue heavy, short sentences). The model had high precision (0.94) but low recall (0.47). Applying a $2.0\times$ scalar boost via the Precision Scalpel improved recall to 0.68 and F1 to 0.79 (+23 points) without degrading precision.
- **Ahmed Shawqi:** Shawqi's archaic vocabulary attracted false positives from various authors. Penalizing his probability class by $0.75\times$ reduced false positives, improving his F1 from 0.84 to 0.92.

6 Conclusion

We presented the system that secured **1st place** in the AbjadAuthorID Shared Task. Our primary finding demonstrates that **morphology outperforms semantics** for Arabic AA: simple character n-grams (0.92 F1) significantly beat fine-tuned AraBERT (0.87 F1), challenging the "Transformer-first" orthodoxy and suggesting that authorial signature in Arabic is encoded more densely in derivational patterns than in semantic topics.

We leveraged this insight to build a hybrid ensemble that uses traditional models for high-precision morphological signal and Deep Learning models for semantic and cross-lingual nuances (such as translation artifacts). By introducing the *Consultant Strategy* (to retain diverse but weaker signals) and the *Precision Scalpel* (for post-hoc calibration), we addressed the challenge of class imbalance and genre confounding, improving recall on difficult authors by up to 23 points.

Future work will focus on automating the calibration process and exploring data augmentation techniques, to address data scarcity for low-resource authors.

7 Limitations

We acknowledge two limitations. First, the *Precision Scalpel* relies on validation statistics, risking overfitting if the test distribution shifts. Second, performance on the lowest-resource author remains capped ($\approx$0.88 F1), suggesting that extreme imbalance (1:9) requires external data augmentation rather than architectural tuning alone.

Acknowledgments

We thank the AbjadAuthorID organizers and for providing datasets for this task.

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.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. Arabert: Transformer-based model for arabic language understanding. *arXiv preprint arXiv:2003.00104*.

John Burrows. 2002. 'delta': a measure of stylistic difference and a guide to likely authorship. *Literary and linguistic computing*, 17(3):267–287.

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.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

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 Implementation Details & Configuration

To ensure full replicability, we report the exact hyperparameters and configuration settings used for our final submission.

A.1 Traditional Model Hyperparameters

We utilized `scikit-learn` for feature extraction and classification. Table 3 details the specific configurations for the three traditional branches.

Parameter	Char	Word	Syntactic
N-Gram Range	(2, 5)	(1, 3)	(3, 5)
Analyzer	char_wb	word	word
Max Features	50,000	30,000	10,000
Sublinear TF	True	True	True
Classifier	LinearSVC		
C	1.0	1.0	0.5
Class Weight	'balanced'		
Calibration	Platt Scaling (cv=3)		

Table 3: Hyperparameters for the three traditional branches. Note the lower regularization ($C = 0.5$) applied to the weaker syntactic branch.

A.2 Deep Learning Configuration

Transformer models were fine-tuned using the HuggingFace `Trainer` API with the following parameters:

- **Models:** `aubmindlab/bert-base-arabertv2` and `xlm-roberta-base`
- **Batch Size:** 8 (Gradient Accumulation = 1)
- **Learning Rate:** 2e-5 (AdamW optimizer)
- **Epochs:** 3 (AraBERT), 4 (XLM-R)
- **Max Length:** 512 tokens (employing a sliding window with stride 256)

A.3 Ensemble Weights

Final ensemble weights (Table 4) were discovered via constrained random search (5,000 iterations), enforcing the constraint $w_{syntactic} \leq 0.15$.

A.4 Precision Scalpel Multipliers

Table 5 lists the class-specific multipliers applied before the final decision. These were empirically tuned on the validation set.

Branch	Weight
Morphological (Char N-Gram)	0.40
Lexical (Word N-Gram)	0.20
Syntactic (POS N-Gram)	0.10
Semantic (AraBERT v2)	0.15
Multilingual (XLM-RoBERTa)	0.15

Table 4: Final voting weights. The morphological signal dominates (40%), while deep learning components contribute 30% combined.

Author	Mult.	Rationale
Tharwat Abaza	2.00×	High Prec (0.94), Low Recall
Kamel Kilani	1.10×	Mild boost for data scarcity
Ahmed Shawqi	0.75×	"Gravity well" (high false pos)
Others	1.00×	No adjustment

Table 5: Calibration multipliers ("Precision Scalpel"). Abaza required a significant boost to overcome genre-based confusion.

B Comprehensive Performance Metrics

B.1 Baseline and Ensemble Comparisons

Table 6 compares individual model performance and costs. Table 7 tracks the incremental improvements during system development.

Model	Val F1	Feat.	Train	Infer
Traditional Models				
Char N-Gram (2-5)	**0.92**	50k	3m	0.1s
Word N-Gram (1-3)	0.89	30k	2m	0.1s
Syntactic (POS 3-5)	0.67	10k	1m	0.1s
Burrows Delta (6 vars)	0.69	300	5m	0.1s
Stylometric Features	0.40	45	1m	<0.1s
Deep Learning Models				
AraBERT v2	0.87	-	2h	5s
XLM-RoBERTa	0.84	-	2.5h	5s

Table 6: Single-Model Baselines. Training time reported for full training set on CPU (Traditional) or P100 GPU (Deep Learning).

B.2 Per-Class Performance

Table 8 details specific improvements for key authors.

C Qualitative Error Analysis

Table 9 highlights the primary confusion patterns discussed in the main text.

D Ablation Visualizations

Figures 1 and 2 illustrate the performance of the baseline models, highlighting the issues (genre confusion and low recall) that motivated our ensemble design.

Config	Val F1	Test F1	Key Component
Char + Word	0.926	-	Best traditional combo
Char alone	0.920	-	Best single model
Unweighted Avg	∼0.91	-	Baseline ensemble
Random Search	0.936	0.920	Weight optimization
+ Constrained	0.936	0.920	Consultant Strategy
+ Pseudo-labeling	0.941	0.930	+3,360 samples
+ Scalpel	**0.948**	**∼0.93**	Calibration

Table 7: Ensemble Configurations showing incremental gains. Pseudo-labeling provided the largest boost to test set generalization (+1.0%).

Author	Base	Final	Gain	N
Tharwat Abaza	**0.56**	**0.79**	**+23**	**90**
Gibran Khalil Gibran	0.88	1.00	+12	30
Ahmed Taymour Pasha	0.81	0.91	+10	57
Ahmed Shawqi	0.84	0.92	+8	171
Kamel Kilani	0.80	0.88	+8	25
Fouad Zakaria	0.85	0.92	+7	132
Naguib Mahfouz	0.95	1.00	+5	327
Hassan Hanafi	0.97	1.00	+3	548
Jurji Zaydan	0.95	0.98	+3	181
Abbas M. al-Aqqad	0.94	0.97	+3	231
Taha Hussein	0.95	0.97	+2	253
Salama Musa	0.93	0.96	+3	142
(Others)	*0.90+*	*0.92+*	*+2-6*	-

Table 8: Per-Class F1 scores comparison. "Base" = Char+Word Baseline, "Final" = Calibrated Ensemble. Tharwat Abaza shows the largest improvement.

True Author	Predicted	Likely Cause
Tharwat Abaza	Wm. Shakespeare	**Genre Overlap:** Both write dramas with short dialogue and imperatives.
Fouad Zakaria	A.G. Makawi	**Topic Overlap:** Both belong to the "Philosopher Cluster" sharing academic vocabulary.
Ahmed Shawqi	Taha Hussein	**Archaic Style:** Shawqi's classical vocabulary acts as a distractor for historical prose.

Table 9: Qualitative error analysis showing primary confusion patterns. Column widths are adjusted to prevent overfull boxes.

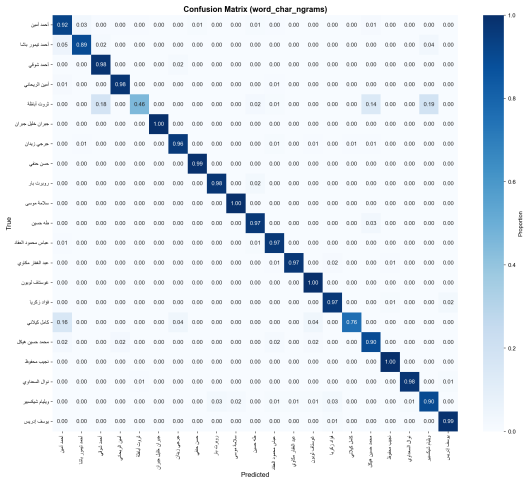

Figure 1: Confusion Matrix of the **N-Gram baseline**. Unlike the final ensemble, this baseline highlights the critical "Genre Confound": note the significant confusion between Tharwat Abaza and translated Shakespeare.

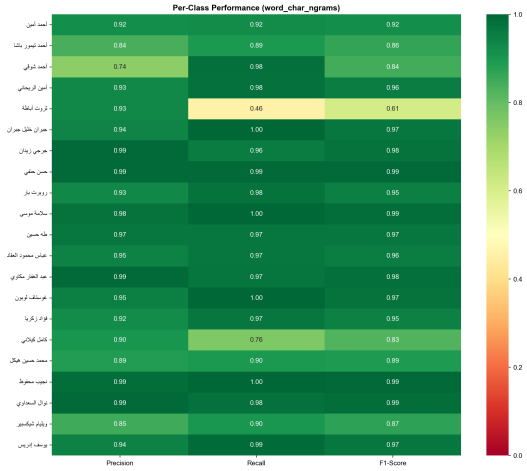

Figure 2: Per-class F1 performance of the **Base Models**. Note the exceptionally low performance on Tharwat Abaza (≈0.46 F1) and reduced precision on Ahmed Shawqi compared to the final system.