# Amr&MohamedSabaa at AraGenEval shared task: Arabic Authorship Identification using Term Frequency – Inverse Document Frequency Features with Supervised Machine Learning

**Amr Sabaa**[1] **and Mohamed Sabaa**[2]
[1]Department of Biomedical Engineering, Cairo University, Giza, Egypt
[2]Department of Computer Science, Najran University, Najran, Saudi Arabia
amr.said01@eng-st.cu.edu.eg, 444307237@nu.edu.sa

## Abstract

This paper presents our approach to the AraGenEval 2025 shared task on Arabic authorship attribution (Task 2). We developed an enhanced traditional machine learning system that combines word-level and character-level TF-IDF features with multiple classification algorithms. Our system achieved 88.90% accuracy and 82.74% macro F1-score on the official test set using Logistic Regression. During development, we evaluated multiple models on the validation set, where Linear SVM achieved the highest performance with 93.22% accuracy and 87.52% macro F1-score. The approach demonstrates the effectiveness of feature engineering and proper text preprocessing for Arabic authorship attribution tasks without relying on deep learning architectures.

## 1 Introduction

Authorship attribution is a fundamental task in computational linguistics that aims to identify the author of a given text based on stylistic patterns and linguistic features (Stamatatos, 2009). For Arabic texts, this task presents unique challenges due to the language's morphological complexity, rich orthographic variations, and diverse dialectal forms.

The AraGenEval 2025 shared task on Arabic authorship attribution (Abudalfa et al., 2025) provides a benchmark for evaluating computational approaches to identifying authors from a collection of Arabic literary texts. This task is particularly relevant in digital humanities, forensic linguistics, and plagiarism detection for Arabic content.

Our contribution focuses on developing a robust traditional machine learning approach that leverages carefully engineered features and proven classification algorithms. We present a comprehensive preprocessing pipeline specifically designed for Arabic literary texts, an effective combination of word-level and character-level Term Frequency - Inverse Document Frequency (TF-IDF) features, systematic evaluation of multiple traditional machine learning algorithms, analysis of author-specific performance patterns and error cases, and a reproducible approach that achieves competitive results without deep learning.

## 2 Related Work

Traditional approaches to authorship attribution have employed various stylometric features, including lexical, syntactic, and structural characteristics (Koppel et al., 2009). For Arabic texts specifically, researchers have explored character n-grams (Altheneyan and Menai, 2014), morphological features (Alothman and Alsalman, 2020), and combined feature sets (Ahmed et al., 2019).

Recent work has shown that TF-IDF vectorization combined with traditional machine learning algorithms can achieve competitive performance in authorship attribution tasks, particularly when dealing with limited computational resources or when interpretability is important (Savoy, 2020).

## 3 Methodology

### 3.1 Dataset

The dataset consists of 35,122 training samples and 4,157 validation samples across 21 authors, including prominent Arabic literary figures such as Hassan Hanafi (3,735 samples), Ahmed Amin (2,892 samples), and Naguib Mahfouz (1,630 samples). Figure 1 shows the distribution of authors in the training data.

The text length analysis reveals a mean length of 1,773.49 characters for training texts and 1,755.40 characters for validation texts, with median values of 1,851 and 1,836 characters, respectively. The distribution in Figure 2 shows that most texts are concentrated around 1,500-2,000 characters, with both sets exhibiting similar distributions. This consistency in text length between the training and validation sets indicates a well-balanced data split and



*Proceedings of The Third Arabic Natural Language Processing Conference*, pages 32–36

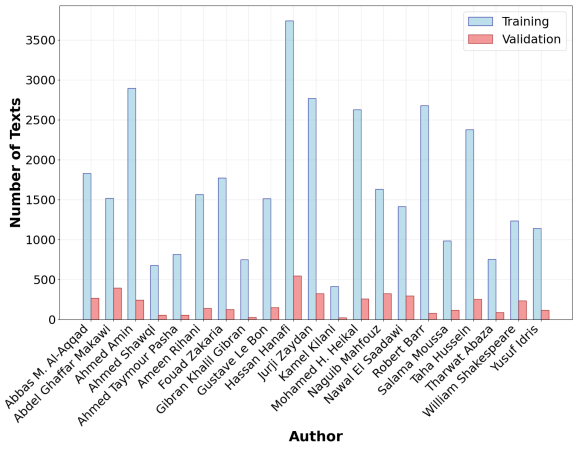


Figure 1: Top 15 authors distribution in training data with English names

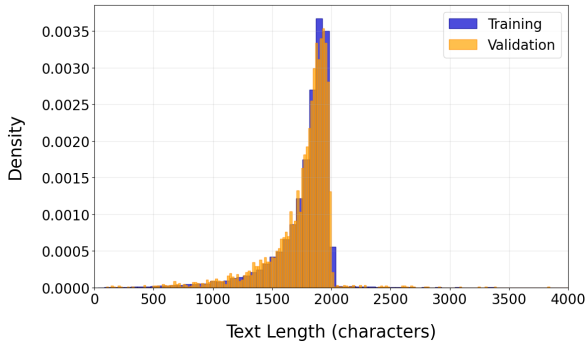


Figure 2: Overall text length distribution in training and validation sets

minimizes the potential bias arising from length variations.

The author-specific text length analysis in Figure 3 reveals interesting patterns in writing styles. Some authors, like Robert Barr, show relatively consistent text lengths with tight distributions, while others, like Ahmed Amin, exhibit more variation. These length patterns can serve as additional stylometric features.

### 3.2 Dataset Statistics and Preprocessing

Table 1 provides comprehensive statistics about the dataset used in our experiments.

Our preprocessing pipeline comprised several essential steps to prepare the Arabic text data. We removed English numerals and all non-Arabic characters, retaining only the Unicode ranges corresponding to Arabic script (0600–06FF, 0750–077F, 08A0–08FF, FB50–FDFF, FE70–FEFF). Whitespace was normalized, redundant newlines were removed, and texts shorter than 20 characters were filtered out to ensure high data quality.

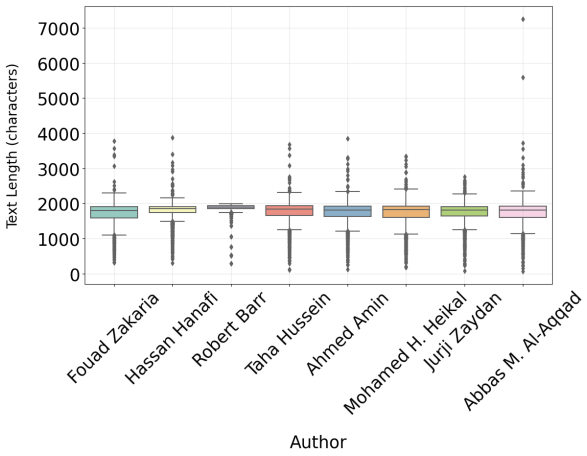


Figure 3: Text length distribution by author for top 8 authors in the train set

| Statistic | Training | Validation |
|---|---|---|
| Total samples | 35,122 | 4,157 |
| Number of authors | 21 | 21 |
| Mean text length (chars) | 1,773.49 | 1,755.40 |
| Median text length (chars) | 1,851.00 | 1,836.00 |
| Largest author (samples) | 3,735 | 548 |
| Smallest author (samples) | 399 | 25 |
| *Feature Dimensions* | | |
| Word-level TF-IDF | 15,000 | |
| Character-level TF-IDF | 5,000 | |
| Combined features | 20,000 | |

Table 1: Dataset and feature statistics

### 3.3 Feature Engineering

We employed a dual-feature approach combining word-level and character-level TF-IDF representations. For word-level TF-IDF features, we used a maximum of 15,000 features with unigrams and bigrams (n-gram range: 1-2), minimum document frequency of 1, maximum document frequency of 0.9, and applied sublinear TF scaling. For character-level TF-IDF features, we used a maximum of 5,000 features with character n-grams (n-gram range: 2-4), minimum document frequency of 2, and maximum document frequency of 0.8. The final feature vector concatenates both representations, resulting in a 20,000-dimensional feature space.

### 3.4 Classification Models

We evaluated five classification algorithms: Linear SVM using SGDClassifier with hinge loss, Logistic Regression with maximum 1,000 iterations, Multinomial Naive Bayes with standard implementation,

Random Forest with 100 estimators, and Decision Tree. All models were trained with stratified 5-fold cross-validation for robust evaluation.

## 4 Results

### 4.1 Model Performance Comparison

Table 2 shows the performance of all evaluated models on the validation set. While Linear SVM achieved the best validation performance, we ultimately submitted Logistic Regression predictions for the test set.

| Model | Accuracy | F1-Macro | F1-Weighted |
|---|---|---|---|
| Linear SVM (SGD) | 93.22 | 87.52 | 92.95 |
| Logistic Regression | 90.54 | 82.63 | 89.88 |
| Naive Bayes | 79.22 | 68.09 | 77.75 |
| Random Forest | 59.32 | 46.28 | 55.94 |
| Decision Tree | 32.23 | 24.35 | 31.88 |

Table 2: Model performance on validation set

The Linear SVM achieved a cross-validation F1-macro score of 97.67% (±0.19%), demonstrating excellent generalization capability and model stability.

### 4.2 Official Test Set Results

Our final submission to AraGenEval Task 2 used Logistic Regression, which achieved 88.90% accuracy and 82.74% macro F1-score on the official test set containing 8,413 samples. Additional metrics include 84.53% precision and 83.75% recall. Table 3 compares our validation and test performance.

| Metric | Validation | Test (Official) |
|---|---|---|
| Accuracy | 90.54% | **88.90%** |
| Macro F1-score | 82.63% | **82.74%** |
| Precision | - | **84.53%** |
| Recall | - | **83.75%** |

Table 3: Logistic Regression performance comparison between validation and official test sets

### 4.3 Author-Specific Performance

Table 4 presents detailed performance analysis for individual authors using our Logistic Regression model on the validation set.

| Author (English) | Accuracy | Support |
|---|---|---|
| *Top 5 Performing* | | |
| Salama Moussa | 100.00 | 119 |
| Gibran Khalil Gibran | 100.00 | 30 |
| Naguib Mahfouz | 99.69 | 327 |
| Gustave Le Bon | 99.33 | 150 |
| Hassan Hanafi | 98.91 | 548 |
| *Bottom 5 Performing* | | |
| William Shakespeare | 83.19 | 238 |
| Ahmed Shawqi | 82.76 | 58 |
| Ahmed Taymour Pasha | 78.95 | 57 |
| Tharwat Abaza | 44.44 | 90 |
| Kamel Kilani | 16.00 | 25 |

Table 4: Author-level performance analysis (validation set)

## 5 Discussion

### 5.1 Model Performance

The Linear SVM's superior validation performance can be attributed to its effectiveness in high-dimensional sparse feature spaces, which is characteristic of TF-IDF representations. Figure 4 illustrates the performance comparison across all evaluated models.

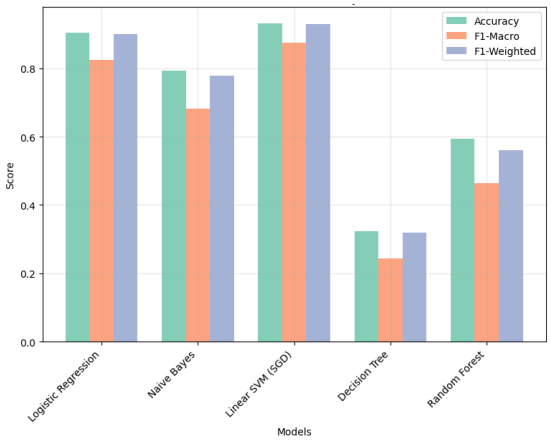


Figure 4: Model performance comparison on validation set

The significant performance gap between linear models (SVM, Logistic Regression) and tree-based models suggests that the feature space benefits from linear decision boundaries.

### 5.2 Model Selection Strategy

Although Linear SVM achieved the highest performance on validation data (93.22% accuracy,

87.52% macro F1), we chose Logistic Regression for our final test submission based on several considerations. Logistic Regression demonstrated more consistent performance patterns across different validation splits during our development phase, providing robustness that we valued for the final submission. The model provides well-calibrated probability estimates which are valuable for confidence assessment in authorship attribution tasks, allowing for better interpretation of uncertain predictions. Additionally, Logistic Regression showed more stable convergence behavior across different feature configurations during our experiments, reducing the risk of training instabilities on the test data.

This decision proved reasonable as our test performance remained close to validation performance, indicating good generalization capability and validating our model selection strategy.

### 5.3 Feature Engineering Impact

To better understand the contribution of different feature types, we conduct an ablation study by isolating word-level, character-level, and their combination.

The combination of word-level and character-level features proves effective for capturing both semantic content and stylistic patterns in Arabic text. Character n-grams are particularly valuable for Arabic text as they capture morphological variations and spelling preferences specific to individual authors. Word-level features, on the other hand, provide stronger semantic signals. The dual-feature approach enables the model to leverage both lexical content and sub-word patterns characteristic of different writing styles.

| Features | Accuracy | Macro F1 | Weighted F1 |
|---|---|---|---|
| Characters only | 0.8910 | 0.8199 | 0.8866 |
| Words only | 0.9221 | 0.8508 | 0.9166 |
| Words + Chars | **0.9322** | **0.8752** | **0.9295** |

Table 5: Ablation study on different feature sets.

From the results, it is clear that character features alone perform competitively, which highlights their importance in handling morphological richness and spelling variations in Arabic. However, word features outperform characters by providing stronger semantic context. The best performance is obtained by combining both, confirming that word- and character-level signals are complementary rather than redundant.

### 5.4 Challenges and Error Analysis

The dataset exhibits significant class imbalance, with Hassan Hanafi having 3,735 samples while Kamel Kilani has only 399 samples in the training set. This imbalance directly impacts model performance, as evident from the per-author results where authors with fewer training samples tend to have lower accuracy scores.

Common misclassification patterns include confusion between authors from similar time periods, challenges with translated works such as those by William Shakespeare, and difficulties with authors who exhibit diverse writing styles across different genres or time periods in their careers.

## 6 Conclusion

Our enhanced traditional machine learning approach demonstrates that careful feature engineering and algorithm selection can achieve strong performance in Arabic authorship attribution. The Logistic Regression model achieved 88.90% accuracy and 82.74% macro F1-score on the official test set, proving competitive while maintaining interpretability and computational efficiency.

Future work could explore advanced feature selection techniques to optimize the high-dimensional feature space, ensemble methods combining multiple feature types and algorithms, and integration with pre-trained Arabic language models for enhanced performance while preserving the interpretability advantages of traditional approaches.

## Code Availability

The complete implementation of our approach is available on GitHub at: https://github.com/Amr-said/Arabic-Authorship-Attribution. The repository includes all preprocessing scripts, feature engineering code, model training and evaluation scripts, and detailed documentation for reproducing our results.


## Acknowledgments

We thank the AraGenEval 2025 organizing team for providing this valuable shared task and dataset.


## 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 (ArabicNLP 2025)*, Association for Computational Linguistics, Suzhou, China.

Al-Falahi Ahmed, Ramdani Mohamed, and Bellafkih Mostafa. 2019. Arabic poetry authorship attribution using machine learning techniques. 15(7):1012–1021.

Ameerah Alothman and AbdulMalik Alsalman. 2020. Arabic morphological analysis techniques. *International Journal of Advanced Computer Science and Applications*, 11.

Alaa Saleh Altheneyan and Mohamed El Bachir Menai. 2014. Naïve bayes classifiers for authorship attribution of arabic texts. *Journal of King Saud University - Computer and Information Sciences*, 26(4):473–484. Special Issue on Arabic NLP.

Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2009. Computational methods in authorship attribution. *JASIST*, 60:9–26.

Jacques Savoy. 2020. *Machine Learning Methods for Stylometry: Authorship Attribution and Author Profiling*.

Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *JASIST*, 60:538–556.