CIOL at AraGenEval shared task: Authorship Identification and AI Generated Text Detection in Arabic using Pretrained Models

Sadia Tasnim Meem and Azmine Toushik Wasi
Computational Intelligence and Operations Laboratory, Bangladesh
Shahjalal University of Science and Technology, Bangladesh
{sadia63,azmine32}@student.sust.edu

Abstract

Authorship identification and AI-generated text detection have recently emerged as pivotal areas of research in natural language processing (NLP), with particular urgency for languages such as Arabic that exhibit complex morphological and orthographic structures. Despite growing interest, most prior work has centered on English and other Indo-European languages, leaving a gap in effective approaches tailored to Arabic's linguistic challenges. This paper presents our participation in two shared tasks: Arabic authorship identification and Arabic AI-generated text detection. For Task2, we fine-tuned transformer-based architectures on a corpus of 21 authors, leveraging parallelized, semantically segmented book data to better capture stylistic variation. For Task3, we trained models on a balanced dataset of human-written and AI-generated news articles produced by multiple large language models. Our approach achieved competitive results across both tasks, underscoring the potential of domain-adapted transformers for morphologically rich languages. We also highlight key limitations, including domain sensitivity and difficulties in distinguishing closely aligned stylistic features, and propose directions for enhancing cross-domain robustness and generalization.

1 Introduction

Authorship identification and AI-generated text detection have emerged as critical research areas in the field of natural language processing (NLP), particularly for languages with complex morphological and orthographic systems such as Arabic. Over the past decade, researchers have developed diverse methodologies for this task, ranging from traditional statistical models to modern deep learning approaches. For instance, ensemble-based strategies have shown promise in enhancing attribution accuracy across heterogeneous datasets (Abbasi et al., 2022). Similarly, deep learning architectures, including convolutional and recurrent neural networks, have been explored for robust authorship identification in multi-domain contexts (Qian et al., 2017). In the domain of Arabic, transformer-based methods such as BERT have been adapted to specific genres, achieving strong results in tasks like poetry authorship attribution (Alqurashi et al., 2025), and knowledge-based models have been utilized to verify authorship in Arabic social media texts (Alqahtani and Yannakoudakis, 2022). Earlier work has also examined fusion approaches for authorship identification in religious Arabic texts, demonstrating the value of multi-feature integration (Sayoud and Hassina, 2021).

Parallel to authorship identification, the increasing sophistication of large language models (LLMs) has introduced the challenge of detecting AI-generated content, especially in morphologically rich languages like Arabic. Recent studies have addressed unique difficulties such as diacritics handling (Alshammari and Elleithy, 2024) and have investigated detection performance in short dialectal Arabic texts (Alharthi, 2025). Encoder-based transformer architectures have also been proposed for Arabic AI-generated text detection, leveraging contextual embeddings for improved accuracy (Alshammari et al., 2024). Comparative evaluations between human and machine-generated Arabic content have further highlighted the challenges of reliably distinguishing AI-authored text from authentic human writing (Boutadjine et al., 2025).

In this paper, we present our systems developed for two shared tasks: (1) Authorship identification in Arabic texts and (2) Arabic AI-generated text detection. We build upon the existing literature in both domains, leveraging transformer-based architectures. Our contributions include fine-tuning domain-specific language models, evaluating their performance on benchmark datasets, and analyzing error patterns to guide future research.

Proceedings of The Third Arabic Natural Language Processing Conference, pages 77–81

2 Background

The shared task (Abudalfa et al., 2025) comprises three subtasks and we worked on two of them: **Task 2** (Authorship Identification) and **Task 3** (Arabic AI-Generated Text Detection). Both are Arabic text classification problems but differ in objectives, input/output formats, and dataset composition.

2.1 Tasks

Task 2: Authorship Identification Task 2 is a multiclass classification problem where the goal is to predict the author of a given text. The input is a paragraph written in the style of a specific author, provided in the `text_in_author_style` column, and the output is the predicted author's name in Arabic, matching the labels in the dataset.

Task 3: Arabic AI-Generated Text Detection Task 3 is a binary classification problem aimed at distinguishing between human-written and AI-generated Arabic news articles or snippets. Human-written samples were sourced from verified news platforms, while AI-generated content was produced using multiple LLMs (e.g., GPT-3.5, GPT-4, Claude) with varied prompting strategies and generation parameters.

2.2 Dataset

For Task 2, the corpus comprises works from 21 authors, each contributing 10 publicly accessible books. Each book was segmented into semantically coherent paragraphs, and selected paragraphs were rephrased into a standardized formal style using GPT-4o mini2, with parallel pairs restricted to at most 1900 tokens. The dataset was split into training, validation, and test sets. For Task 3, the dataset contains human-written content sourced from verified news platforms and AI-generated content produced by multiple LLMs (e.g., GPT-3.5, GPT-4, Claude) under varied prompting strategies and generation parameters. It includes 4,800 training samples, a forthcoming development set, and 2,000 test samples, with a balanced distribution of human and AI-generated texts.

3 System Overview

This section outlines the architectures and strategies employed in our system for the shared tasks.

3.1 Task 2: Authorship Identification

In this subsection, we describe our approach to modeling authorial style and capturing distinctive linguistic features for the authorship identification task.

Key Algorithms and Design Decisions. For Task 2, we adopted the `CAMeL-Lab/bert-base-arabic-camelbert-mix` pretrained language model due to its strong performance on Arabic text understanding and ability to capture fine-grained stylistic differences critical for authorship attribution. The task was framed as a *multiclass classification* problem over $N = 21$ authors. Each paragraph was tokenized to a maximum length of 512 tokens with dynamic padding. The BERT classification head was replaced with a dense layer of size N, followed by softmax. The model was fine-tuned end-to-end using cross-entropy loss.

Addressing Task Challenges. The authorship identification task presented several challenges. First, many authors exhibited highly similar writing styles, making stylistic differentiation difficult; this was mitigated through the use of contextualized embeddings from the pretrained transformer, which capture subtle variations in style. Second, the dataset contained long paragraphs, often exceeding the model's input length; to address this, we truncated inputs to 512 tokens while prioritizing semantically important segments to preserve representative style cues. Finally, although class imbalance was relatively minor, it still posed risks of skewed evaluation, so we did not apply resampling but instead relied on macro-F1 as the primary metric to ensure fairness across authors. These design choices collectively allowed the model to handle the practical difficulties of morphologically rich Arabic text while maintaining robust performance.

System Configuration. Training was conducted for 4 epochs using the AdamW optimizer with a learning rate of 2×10^{-5}, batch size of 16, and weight decay of 0.01. Model selection was performed based on the highest validation macro-F1 score to ensure balanced performance across all author classes. Evaluation metrics included both accuracy, to capture overall correctness, and macro-F1, to account for class imbalance and provide a fairer assessment of performance across authors.

3.2 Task 3: Arabic AI-generated Text Detection

Here, we present our methodology for distinguishing between human-written and AI-generated Arabic text across multiple domains.

3.2.1 Configuration 1

We used `AraBERTv2`[1] for binary classification of human-written (1) versus machine-generated (0) text. The preprocessing stage involved mapping labels, replacing missing entries with empty strings, and applying a stratified train–validation split to handle class imbalance. Text was tokenized with the `AraBERTv2` tokenizer using a maximum sequence length of 512 tokens. The model consisted of the pretrained `aubmindlab/bert-base-arabertv2` encoder, followed by dropout ($p = 0.3$), a dense layer with two output units, and a softmax classifier. Training was performed with cross-entropy loss, gradient clipping ($\|g\|_\infty \leq 1.0$), and early stopping to prevent overfitting, ensuring robust performance on Arabic-specific tokenization challenges.

3.2.2 Configuration 2

In this variant, we employed `aubmindlab/bert-base-arabert` with `AutoModelForSequenceClassification`, which simplified implementation by providing a built-in classification head. Tokenization was limited to a maximum length of 256 tokens to improve efficiency and reduce memory usage. The model consisted of the BERT encoder paired with the classification head for two output classes, trained using the AdamW optimizer with a linear learning rate scheduler over 3 epochs. Pretrained weights from `aubmindlab/bert-base-arabert` were used to leverage prior Arabic language knowledge. While the shorter sequence length improved computational efficiency, it slightly impacted performance; model evaluation was monitored using accuracy, precision, recall, and F1 to ensure balanced assessment across metrics.

For Task 3, Configuration 1 outperformed Configuration 2 due to longer context handling, stronger pretrained embeddings, and custom classifier design.

4 Experimental Setup

4.1 Dataset Processing

For both tasks, the datasets were divided into training, development, and test sets as provided. The training sets were used to train the models, the development sets for validation and hyperparameter tuning, and the test sets for final evaluation. For Task 2, the official training and development sets were used, while for Task 3, training was performed on the provided files and evaluation was done on the official unlabelled file.

4.2 Preprocessing and Hyperparameter Details

Text preprocessing included Arabic-specific normalization, removal of non-Arabic characters, and lowercasing to promote uniformity across inputs. Tokenization was performed using the `AutoTokenizer` from Hugging Face Transformers, with a maximum sequence length of 256 tokens for Task 2 and 512 tokens for Task 3, reflecting the different input requirements of each task. Training batch sizes were set to 16 for Task 2 and 8 for Task 3. Models were optimized using the AdamW optimizer with a learning rate of 2×10^{-5}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1 \times 10^{-8}$, along with a linear learning rate warmup over 10% of the total training steps. Task 2 models were trained for 4 epochs, while Task 3 models were trained for 3 epochs. Dropout layers and gradient clipping were applied as described in the system section to prevent overfitting and stabilize training, ensuring consistent convergence across different runs and input variations.

4.3 Evaluation Metrics

Model performance was evaluated using accuracy and F1 metrics. For Task 2, macro-F1 was used to account for class imbalance across the 21 authors, with accuracy as a complementary measure. For Task 3, F1 and accuracy were employed to capture both the balance between precision and recall and overall correctness.

5 Results

5.1 Task 2: Authorship Identification

Evaluation Set Results. We evaluated the fine-tuned CAMeL-BERT model on the development and test splits. On the held-out validation set, the model achieved a final evaluation loss of 0.584, accuracy of 0.872, and macro-F1 score of 0.809 after 4 epochs. Table 1 shows the epoch-wise training and validation metrics.

Test Set Results. For the final test submission, the model achieved an F1-score of 0.827, accuracy of 0.864, precision of 0.828, recall of 0.854, specificity of 0.854, and balanced accuracy of 0.854. The system ranked competitively among all submissions.

[1] https://huggingface.co/aubmindlab/bert-base-arabertv2

Table 1: Task 2: Epoch-wise training results on the validation set

Epoch	Training Loss	Validation Loss	Accuracy	F1
1	0.1655	0.6413	0.8273	0.7478
2	0.0591	0.5431	0.8595	0.7774
3	0.0055	0.6400	0.8643	0.7995
4	0.0145	0.5842	0.8723	0.8093

Quantitative Findings and Analysis. Comparing epoch-wise development set performance and test submission results, we observe that the design choices—such as stratified splitting, 512-token input length, and dropout regularization—contributed positively to overall generalization. Ablation of dropout or reducing sequence length to 256 tokens led to a drop in macro-F1 by 2–3% on validation. Using CAMeL-BERT's contextual embeddings for Arabic significantly improved performance compared to simpler baselines such as TF-IDF + Logistic Regression (macro-F1 $\sim$0.65).

5.2 Task 3: Arabic AI-Generated Text Detection

Evaluation Set Results. For Task 3, we experimented with two approaches for detecting AI-generated Arabic text. The approach that performed better was selected for detailed reporting. On the held-out validation set the model was trained for 3 epochs and achieved the following performance. On the held-out validation set, the model achieved a validation loss of 0.0861, an accuracy of 0.9844, an F1-score of 0.9841, a precision of 1.0000, and a recall of 0.9688. Epoch-wise training results are summarized in Table 2.

Table 2: Task 3: Epoch-wise training results on the validation set

Epoch	Training Loss	Validation Loss	Accuracy	F1
1	0.1013	0.1271	0.9781	0.9777
2	0.0197	0.0564	0.9896	0.9895
3	0.0047	0.0861	0.9844	0.9841

Test Set Results. On the official test split, the selected model achieved an F1-score of 0.657, an accuracy of 0.704, a precision of 0.780, a recall of 0.568, a specificity of 0.840, and a balanced accuracy of 0.704.

Quantitative Findings and Analysis. Although the validation performance was very high (F1 $\sim$0.984), the official test results indicate a substantial drop in F1-score (0.657) and recall (0.568). This suggests a significant domain shift between the training/validation data and the test data or the presence of challenging AI-generated text patterns not seen during training. The high precision (0.780) and specificity (0.840) indicate that the model is conservative in predicting AI-generated text, favoring fewer false positives but missing a considerable portion of AI-generated instances.

Overall, the results highlight that while contextual embeddings and fine-tuning strategies can achieve near-perfect validation performance, careful attention to dataset diversity and robustness is necessary for generalization to unseen test examples. Future work should consider data augmentation, cross-domain evaluation, and adversarial training to better detect AI-generated Arabic text.

6 Conclusion

In this study, we have presented systems for two Arabic NLP tasks: authorship identification (Task 2) and AI-generated text detection (Task 3). For Task 2, a fine-tuned CAMeL-BERT model achieved strong performance, with 87% accuracy and a macro-F1 score of 0.809 on the validation set, demonstrating its ability to effectively capture and model distinctive authorial styles in a morphologically rich language like Arabic. Task 3 employed a contextual embedding-based approach for distinguishing human-written from AI-generated text, achieving near-perfect performance on the validation set (F1 $\sim$0.984). However, the official test results showed a notable drop (F1 = 0.657), highlighting the challenges of generalizing to unseen AI-generated content and the variability introduced by different text sources and generation methods. These findings emphasize the importance of domain adaptation and robust evaluation strategies when deploying NLP models for Arabic text analysis.

Overall, our results demonstrate the promise of transformer-based models for both stylistic and generative text classification tasks, while also underlining the need for further research on cross-domain generalization and handling the evolving capabilities of large language models.

Limitations

Despite achieving strong performance, our study has several limitations. In Task 2, distinguishing authors with subtle stylistic differences remains challenging, particularly when writing styles overlap or when texts are short. For Task 3, AI-generated text detection proved sensitive to domain shifts, resulting in reduced generalization to unseen sources or generation methods. Future work should investigate more advanced transformer-based architectures, data augmentation techniques, and cross-domain training to enhance robustness. Additionally, incorporating explainable AI methods could provide greater transparency and interpretability of model decisions. Beyond technical considerations, these findings have broader implications: improving authorship identification and AI-generated content detection in Arabic can support academic integrity, media verification, and responsible AI deployment, helping to mitigate the spread of misinformation and enhance trust in digital content.

Broader Impact Statement

The development of robust authorship identification and AI-generated text detection systems for Arabic has important societal implications. These tools can help maintain academic integrity by detecting plagiarism, support media and news verification to combat misinformation, and promote responsible use of AI-generated content. Moreover, advancing NLP methods for morphologically rich languages like Arabic contributes to more inclusive AI technologies, ensuring that non-English languages benefit from state-of-the-art models and reducing linguistic biases in automated text analysis. By improving transparency and accountability in content generation and evaluation, such systems can foster trust in digital communication and AI applications more broadly.

References

Ahmed Abbasi, Asad R. Javed, Fahad Iqbal, and 1 others. 2022. Authorship identification using ensemble learning. *Scientific Reports*, 12:9537.

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)*, Suzhou, China. Association for Computational Linguistics.

Haifa Alharthi. 2025. Investigation into the identification of ai-generated short dialectal arabic texts. *IEEE Access*, PP:1–1.

Fatimah Alqahtani and Helen Yannakoudakis. 2022. Authorship verification for arabic short texts using arabic knowledge-base model (arakb). In *Proceedings of the Seventh Arabic Natural Language Processing Workshop (WANLP)*, pages 205–213.

Lama Alqurashi, Serge Sharoff, Janet Watson, and Jacob Blakesley. 2025. BERT-based classical Arabic poetry authorship attribution. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6105–6119, Abu Dhabi, UAE. Association for Computational Linguistics.

H. Alshammari, A. El-Sayed, and K. Elleithy. 2024. Ai-generated text detector for arabic language using encoder-based transformer architecture. *Big Data and Cognitive Computing*, 8(3):32.

H. Alshammari and K. Elleithy. 2024. Toward robust arabic ai-generated text detection: Tackling diacritics challenges. *Information*, 15(7):419.

Amal Boutadjine, Fouzi Harrag, and Khaled Shaalan. 2025. Human vs. machine: A comparative study on the detection of ai-generated content. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 24(2).

Chen Qian, Tianchang He, and Rao Zhang. 2017. Deep learning based authorship identification. *Report, Stanford University*, pages 1–9.

H. Sayoud and Hadjadj Hassina. 2021. Authorship identification of seven arabic religious books -a fusion approach. *The Journal of Scientific and Engineering Research*, 6:137–157.