PTUK-HULAT at AraGenEval Shared Task: Fine-Tuning XLM-RoBERTa for AI-Generated Arabic News Detection

Tasneem Duridi
Computer Science Department
Palestine Technical University - kadoorie
Tulkarm, Palestine
tasneem.duridi@ptuk.edu.ps

Areej Jaber
Computer Science Department
Palestine Technical University - kadoorie
Tulkarm, Palestine
a.jabir@ptuk.edu.ps

Paloma Martínez
Computer Science Department
Universidad Carlos III de Madrid
Madrid, Spain
pmf@inf.uc3m.es

Abstract

The authenticity of digital content has become an increasingly critical challenge with the rapid adoption of generative AI tools, especially for low-resource languages such as Arabic. The language's rich morphology and domain diversity further complicate the detection of machine-generated Arabic text. In this work, we present our submission to the ARATECT 4.3 shared task, Subtask 3, which focuses on identifying AI-generated Arabic news articles. Our approach employs fine-tuned multilingual transformer models based on XLM-RoBERTa. The XLM-RoBERTa-large model achieved a macro F1-score of 0.93 on the development set, while the XLM-RoBERTa-base model obtained an F1-score of 0.78 on the test set, ranking fourth on the official leaderboard. This paper outlines our methodology, presents the experimental results, and discusses key insights from our participation.

1 Introduction

The rapid development of large language models (LLMs), such as GPT-4 (Achiam et al., 2023), PaLM (Chowdhery et al., 2023), and ChatGPT (Maniaci et al., 2024), has enabled the generation of coherent and contextually rich text from simple prompts. These models have transformed natural language generation (NLG), supporting applications in education, journalism, scientific writing, and customer service (Duaibes et al., 2024). However, their widespread adoption has also raised concerns regarding the authenticity and ethical implications of AI-generated text (AIGT), particularly in high-stakes domains (Stahl and Eke, 2024; Cotton et al., 2024).

Distinguishing AIGT from human-written text (HWT) remains a persistent challenge, especially as modern systems such as ChatGPT and Gemini (Imran and Almusharraf, 2024) increasingly emulate natural human language. Misuse of such technology has been associated with misinformation, plagiarism, and declining trust in online content (Weidinger et al., 2022; Sheng et al., 2021; Gao et al., 2022; Duridi et al., 2025; Jazzar and Duridi, 2024). Despite efforts to develop detection tools, most are designed for English or other Latin-script languages, with limited adaptation for morphologically rich, low-resource languages.

Arabic, spoken by over 440 million people across 22 countries (Jaber and Martínez, 2023), remains underrepresented in AIGT detection research. Its complex morphology, optional diacritics, and stylistic diversity present unique challenges for existing detection systems (Duridi et al., 2024). Only a few recent studies have directly addressed Arabic AIGT detection (Alshammari et al., 2024), and some report performance degradation when models are applied to diacritized Arabic HWT (Alshammari and Ahmed, 2023).

To address this gap, the AraGenEval Shared Task introduced ARATECT Subtask 3: Arabic News Text Detection (Abudalfa et al., 2025), which focuses on distinguishing human-written from AI-generated Arabic news articles. For this subtask, the PTUK-HULAT team developed a detection system based on multilingual transformer models fine-tuned on stratified splits of the shared task dataset. Our primary system, built on XLM-RoBERTa-base, achieved an F1-score of 0.78 on the test set, ranking fourth on the official leaderboard. The implementation code is publicly available at: GitHub Repository.

Proceedings of The Third Arabic Natural Language Processing Conference, pages 42–48

2 Background

ArabicNLP 2025 features eleven shared tasks, including Shared Task 5: AraGenEval on Arabic Authorship Style Transfer (AST) and AI-Generated Text (AIGT) detection. Within this task, ARATECT 4.3 (Abudalfa et al., 2025) evaluates systems on distinguishing between human-written and AI-generated Arabic text across multiple genres. Subtask 3 — Arabic News Text Detection (ArabicNewsGen) — focuses on classifying full-length Arabic news articles and shorter excerpts into two categories: human-written or AI-generated.

The input to the system consists of a single Arabic news text, which may range from short passages to full-length articles. The output is a binary label: `human` for human-written or `machine` for AI-generated. Table 4 provides representative examples from each class in Appendix A.

3 Related Work

Research on AIGT detection has largely focused on English, with early tools like GPTZero and OpenAI's classifier targeting synthetic content. The rise of Arabic generative models has prompted studies on Arabic-specific detection methods.

(Antoun et al., 2020b) introduced AraGPT2 alongside a discriminator trained to detect its outputs, achieving up to 98% accuracy. They later developed AraELECTRA (Antoun et al., 2020a), an Arabic adaptation of ELECTRA (Clark et al., 2020), which demonstrated strong performance in distinguishing real from synthetic Arabic texts. Harrag et al. (Harrag et al., 2021) fine-tuned AraBERT on synthetic Arabic tweets, outperforming traditional sequence models with 98.7% accuracy. Other studies (Almerekhi and Elsayed, 2015; Alghamdi and Alowibdi, 2024) applied classical machine learning with handcrafted features to detect bot-generated Arabic social media content, reporting around 92% accuracy.

More recent work by Alshammari et al. (Alshammari and Ahmed, 2023) highlighted the limitations of general-purpose detectors for Arabic, proposing fine-tuned AraELECTRA and XLM-RoBERTa models on ChatGPT- and Bard-generated datasets, achieving near 99% accuracy after dediacritization. Alharthi (Alharthi, 2025) addressed detection in multiple Arabic dialects, providing novel dialectal datasets and achieving up to 97% accuracy with fine-tuned AraELECTRA and AraBERT, emphasizing the challenge of paraphrased content and the importance of features like lexical diversity and readability.

These studies illustrate the progress and ongoing challenges in Arabic AIGT detection, particularly the need for dialect-aware datasets, robust benchmarks, and models capable of cross-dialect generalization.

4 Dataset

The organizers of the ArabicNewsGen shared task released a dataset containing Arabic news articles in various domains, including politics, economy, technology and sports, and was released in three phases, as summarized in Table 1. The training set contains 4,798 labeled articles (`id`, `content`, `label`), moderately balanced across the `human` and `machine` classes; approximately 1.3% of entries with missing `content` were removed during preprocessing. The development set consists of 500 unlabeled articles (`id`, `title`, `content`) for validation and tuning, while the test set includes 500 unlabeled articles with the same structure as the development set, used for leaderboard-based evaluation against hidden labels.

5 System Description

Our model selection process was iterative. We began by fine-tuning several widely used Arabic and multilingual transformers, including mBERT, DistilBERT, QARiBERT, and AraELECTRA. Among these, AraELECTRA achieved the highest score on the test set. Although mBERT, DistilBERT, and QARiBERT produced relatively strong results during training, AraELECTRA and XLM-RoBERTa consistently delivered stronger and more reliable performance across both the development and test sets. This finding aligns with prior studies (see Section 3), which highlight AraELECTRA's effectiveness in Arabic-specific tasks and XLM-RoBERTa's robustness in handling multilingual and mixed-language text. Based on these observations, we prioritized AraELECTRA and XLM-RoBERTa (base and large) in our final evaluation, along with a BiLSTM-enhanced variant of XLM-RoBERTa-base.

5.1 Models

AraELECTRA is an Arabic-specific model based on the ELECTRA framework (Antoun et al., 2020a), which uses a replaced token detection pretraining objective. Pre-trained solely on exten-

Table 1: Summary of the ARATECT 4.3 Subtask 3 dataset.

Phase	Samples	Fields	Avg Length (words)	English (%)
Training	4,798	id, content, label	485.77	15.86
Development	500	id, title, content	288.74	56.60
Testing	500	id, title, content	238.96	37.60

sive Arabic corpora, AraELECTRA offers efficient training and strong performance on Arabic NLP tasks, making it well-suited for AI-generated text detection in Arabic news domains.

XLM-RoBERTa-base and XLM-RoBERTa-large are multilingual transformer models trained on 2.5TB of CommonCrawl data across 100 languages (Conneau et al., 2019). The base model contains 270 million parameters, providing a balance between performance and computational efficiency, while the large model scales up to 550 million parameters to capture richer linguistic patterns.

XLM-RoBERTa-base + BiLSTM extends the base transformer by adding a BiLSTM layer atop the transformer encoder outputs to model sequential dependencies and stylistic flow more effectively. The BiLSTM processes the summed embeddings from the last four transformer layers bidirectionally, enabling the capture of long-range contextual patterns indicative of AI-generated text. During fine-tuning, only the last four transformer layers are unfrozen to maintain pre-trained knowledge, while the BiLSTM and classifier layers are trained fully. The BiLSTM hidden size is set to 256 units with a single bidirectional layer.

6 Experimental Setup

6.1 Data and Preprocessing

We utilized the provided labeled dataset, splitting it into training (90%) and development (10%) subsets using stratified sampling to preserve class distributions.

Preprocessing involved removing samples with empty content fields and concatenating the title and content fields into a single text sequence. The textual class labels (human and machine) were mapped to numerical labels, with human assigned 0 and machine assigned 1.

Although we initially experimented with extensive text cleaning—including removing diacritics, normalizing Arabic letters, eliminating punctuation, and collapsing repeated characters—we observed that applying these steps actually reduced model performance. Therefore, no additional text cleaning or normalization was applied prior to tokenization, as keeping the raw text produced better results.

6.2 Training Details

All models were trained using the AdamW optimizer (Loshchilov and Hutter, 2017) with early stopping (patience=3 epochs) based on the development set F1 score for the machine class. Hyperparameters were selected through empirical validation considering model architecture and size constraints.

Key hyperparameter ranges across experiments:

- Learning rate: 110^{-5} to 510^{-5}
- Batch size: 4-16 (adjusted for model memory requirements)
- Dropout: 0.1-0.5 (higher for more complex architectures)
- Warmup ratio: 0-10% of total training steps
- Label smoothing: $\epsilon = 0.0 - 0.1$
- Maximum epochs: 10-20

For consistency across experiments, we employed weighted random sampling and class-weighted cross-entropy loss in all training runs, though the training data was balanced. The specific hyperparameter configurations for each model variant are provided in Table 5 in Appendix B.

6.3 Implementation and Evaluation

Experiments were run on Google Colab with NVIDIA T4 GPUs, leveraging PyTorch, HuggingFace Transformers, and the Accelerate library for efficient training. Evaluation metrics included precision, recall, and F1-score per class.

7 Results

7.1 Development Phase Performance

Table 2 demonstrates the superior performance of XLM-RoBERTa-large on the development set,

achieving state-of-the-art results with 0.9272 F1-score and 92.4% accuracy. The model exhibits exceptional recall (0.968), indicating near-perfect detection of machine-generated texts. While XLM-RoBERTa-base shows solid performance (0.8532 F1), AraELECTRA's high recall (0.912) is offset by low precision (0.5078), revealing language-specific challenges in Arabic AIGT detection and limiting its suitability for further evaluation.

7.2 Test Phase Performance

On the test set Table 3, XLM-RoBERTa-base maintains the strongest balance between precision and recall (0.7823 F1). The BiLSTM-enhanced variant shows a distinct precision-focused profile (0.8029 precision vs. 0.668 recall), suggesting architectural modifications significantly impact error tradeoffs. Performance degradation from development to test sets (XLM-R-base F1: 0.8532 $\rightarrow$ 0.7823) highlights domain shift challenges in AIGT detection.

The experimental results demonstrate that the XLM-RoBERTa-large model significantly outperforms the base variant on the development set, benefiting from its enhanced capacity to capture the complex linguistic patterns necessary for distinguishing between human- and machine-generated Arabic texts. The model's high recall and balanced accuracy indicate its effectiveness in identifying machine-generated content, which is critical for practical detection applications.

On the test set, the XLM-RoBERTa-base model achieves a more balanced trade-off between recall and precision compared to the BiLSTM-enhanced variant. While the BiLSTM addition improves precision and specificity, it does so at the expense of recall, resulting in a more conservative classifier that may fail to detect certain machine-generated samples. This trade-off underscores the need to carefully select model architectures according to the intended application's prioritization of recall versus precision.

The inherent characteristics of the dataset—such as predominantly Arabic text with a minor English component, variable text lengths, and the presence of abbreviations—pose challenges that larger transformer-based models are often better equipped to address due to their richer representational capacity. Furthermore, differences in text length and language composition between the training and evaluation sets likely contribute to domain shifts, which may explain the observed performance degradation on the test set relative to development results. Not all models from the development phase were carried forward to the test phase: AraELECTRA, despite its high recall, exhibited poor precision and overall F1-score, making it unreliable for balanced AIGT detection. XLM-RoBERTa-large achieved the best performance on the development set, but its evaluation on the test set was excluded due to substantial computational cost. Therefore, the test set experiments focused on XLM-RoBERTa-base and its BiLSTM-enhanced variant, which offered a practical balance between efficiency and performance while allowing exploration of architectural improvements.

8 Conclusion

This work investigated multiple transformer-based architectures for detecting AI-generated Arabic text, including XLM-RoBERTa-base, XLM-RoBERTa-large, and a BiLSTM-enhanced variant. The best development set performance was achieved by XLM-RoBERTa-large, benefiting from its higher representational capacity to capture complex Arabic linguistic patterns. On the test set, XLM-RoBERTa-base offered a more balanced precision–recall trade-off, while the BiLSTM addition improved specificity at the cost of recall.

Despite strong results, the system faces challenges from domain shifts between training and test data, varying text lengths, and mixed-language content, which reduce performance on unseen data. Future work will address these issues through domain adaptation, better model designs for balancing precision and recall, and improvements to handle diverse Arabic texts and code-switching.

Acknowledgments

This work has been supported by the Palestine Technical University- Khadoori and Universidad Carlos III de Madrid.

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.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman,

Table 2: Development Set Performance Comparison

Model	F1-score	Precision	Recall	Accuracy
XLM-RoBERTa-large	0.9272	0.8897	0.9680	0.9240
XLM-RoBERTa-base	0.8532	0.8352	0.8720	0.8500
AraELECTRA	0.6524	0.5078	0.9120	0.5140

Table 3: Test Set Performance Comparison

Model	F1-score	Precision	Recall	Accuracy
XLM-RoBERTa-base	0.7823	0.7260	0.8480	0.7640
XLM-RoBERTa + BiLSTM	0.7293	0.8029	0.6680	0.7520

Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Noura Saad Alghamdi and Jalal Suliman Alowibdi. 2024. Distinguishing arabic genai-generated tweets and human tweets utilizing machine learning. *Engineering, Technology & Applied Science Research*, 14(5):16720–16726.

Haifa Alharthi. 2025. Investigation into the identification of ai-generated short dialectal arabic texts. *IEEE Access*.

Hind Almerekhi and Tamer Elsayed. 2015. Detecting automatically-generated arabic tweets. In *AIRS*, pages 123–134. Springer.

Hamed Alshammari and El-Sayed Ahmed. 2023. Airabic: Arabic dataset for performance evaluation of ai detectors. In *2023 International Conference on Machine Learning and Applications (ICMLA)*, pages 864–870. IEEE.

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

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020a. Araelectra: Pre-training text discriminators for arabic language understanding. *arXiv preprint arXiv:2012.15516*.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020b. Aragpt2: Pre-trained transformer for arabic language generation. *arXiv preprint arXiv:2012.15520*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, and 1 others. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

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*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Debby RE Cotton, Peter A Cotton, and J Reuben Shipway. 2024. Chatting and cheating: Ensuring academic integrity in the era of chatgpt. *Innovations in education and teaching international*, 61(2):228–239.

Lina Duaibes, Areej Jaber, Mustafa Jarrar, Ahmad Qadi, and Mais Qandeel. 2024. Sina at fignews 2024: Multilingual datasets annotated with bias and propaganda. *arXiv preprint arXiv:2407.09327*.

Tasneem Duridi, Lour Atwe, Areej Jaber, Eman Daraghmi, and Paloma Martínez. 2025. Detection of propaganda and bias in social media: A case study of the israel-gaza war (2023). In *2025 International Conference on New Trends in Computing Sciences (ICTCS)*, pages 204–210. IEEE.

Tasneem Duridi, Derar Eleyan, Amna Eleyan, and Tarek Bejaoui. 2024. Arabic fake news detection using machine learning approach. In *2024 International Symposium on Networks, Computers and Communications (ISNCC)*, pages 1–7. IEEE.

Catherine A Gao, Frederick M Howard, Nikolay S Markov, Emma C Dyer, Siddhi Ramesh, Yuan Luo, and Alexander T Pearson. 2022. Comparing scientific abstracts generated by chatgpt to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers. *BioRxiv*, pages 2022–12.

Fouzi Harrag, Maria Debbah, Kareem Darwish, and Ahmed Abdelali. 2021. Bert transformer model for detecting arabic gpt2 auto-generated tweets. *arXiv preprint arXiv:2101.09345*.

Muhammad Imran and Norah Almusharraf. 2024. Google gemini as a next generation ai educational tool: a review of emerging educational technology. *Smart Learning Environments*, 11(1):22.

Areej Jaber and Paloma Martínez. 2023. Ptuk-hulat at araieval shared task fine-tuned distilbert to predict disinformative tweets. In *Proceedings of ArabicNLP 2023*, pages 525–529.

Mahmoud Jazzar and Tasneem Duridi. 2024. A comprehensive review of machine learning and deep learning techniques for cyberbullying detection. In *International Conference on Smart Cyber Physical Systems*, pages 1–12. Springer.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Antonino Maniaci, Carlos M Chiesa-Estomba, and Jérôme R Lechien. 2024. Chatgpt-4 consistency in interpreting laryngeal clinical images of common lesions and disorders. *Otolaryngology–Head and Neck Surgery*, 171(4):1106–1113.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2021. Societal biases in language generation: Progress and challenges. *arXiv preprint arXiv:2105.04054*.

Bernd Carsten Stahl and Damian Eke. 2024. The ethics of chatgpt–exploring the ethical issues of an emerging technology. *International Journal of Information Management*, 74:102700.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, and 1 others. 2022. Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM conference on fairness, accountability, and transparency*, pages 214–229.

A Appendix A: Sample Arabic Texts

B Appendix B: Key Training Hyperparameters

Table 4: Sample Arabic Texts with Labels

Content	Label
ذكر تقرير لمجلة فوربس أن عمليات الاحتيال الإلكتروني عند السفر -بما في ذلك سرقة الهوية والاحتيال المصرفي والاحتيال باستخدام بطاقات الائتمان- تشهد تصاعداً ملحوظاً مع تقدم التكنولوجيا وظهور تقنيات مثل الذكاء الاصطناعي التي تُستخدم لتطوير هجمات أكثر تعقيداً. ووفقاً المجلة شهدت عمليات الاحتيال المرتبطة بالسفر زيادة كبيرة -كما أوردت مركز موارد سرقة الهوية آي تي آر سي يةضرر- وهو ما يسلط الضوء على ضرورة اتخاذ تدابير لحماية البيانات الشخصية والمالية. بحسب تقرير فوربس لتعزيز الحماية الإلكترونية وأكد تقرير فوربس أن التطور السريع في تقنيات الاحتيال الإلكتروني يتطلب يقظة دائمة من المستهلكين، ناصحا بعدم مشاركة المعلومات إلا مع جهات موثوقة ومحذرا من التعامل مع أي تواصل إلكتروني غير موثوق.	human
وصل عدد من ضحايا الغارات الإسرائيلية في قطاع غزة إلى المستشفى المعمداني لتلقي العلاج الضروري. تعرض الضحايا لإصابات خطيرة نتيجة الهجمات الجوية الأخيرة التي نفذتها إسرائيل في المنطقة. يستمر العاملون في المستشفى في تقديم الرعاية الطبية اللازمة للمصابين والعمل على استقرار حالتهم الصحية.	machine

Table 5: Key training hyperparameters per model architecture

Parameter	XLM-R Base	BiLSTM	XLM-R Large	Arabic ELECTRA
Learning rate	310^{-5}	510^{-5}	310^{-5}	310^{-5}
Batch size	16	16	4	16
Max epochs	10	20	10	10
Warmup ratio	10%	0%	10%	10%
Dropout	0.1	0.5	0.1	0.1
Label smoothing (ϵ)	0.1	-	0.1	0.1
Optimizer	AdamW			
Early stopping	Patience=3 (F1)			
Class weighting	Yes			