LMSA at AraGenEval Shared Task: Ensemble-Based Detection of AI-Generated Arabic Text Using Multilingual and Arabic-Specific Models

Kaoutar Zita[1]*, Attia Nehar[2], Abdelkader Khelil[2], Slimane Bellaouar[1], Hadda Cherroun[3]

[1]Laboratoire des Mathématiques et Sciences Appliquées (LMSA), Université de Ghardaia, Algeria
[2]Faculty of Exact Sciences and Computer Science, University of Djelfa, Algeria
[3]Laboratoire d'informatique et des Mathématiques, Université Amar Telidji, Laghouat, Algeria
{zita.kaoutar, bellaouar.slimaneg}@univ-ghardaia.edu.dz,
{neharattia, a.khelil}@univ-djelfa.dz,
hadda.cherroun@lagh-univ.dz

Abstract

We address the problem of distinguishing between human-authored and AI-generated text in low-resource languages, particularly Arabic. We present the LMSA[1] team's participation in the ARATECT (Arabic AI-Generated Text Detection) subtask of the AraGenEval[2] shared task, which targets the detection of AI-generated Arabic texts. We propose an ensemble-based classification framework that integrates multilingual and Arabic-specific pre-trained language models, namely Fanar, AraBERT, and XLM-R, optimized through a dedicated fine-tuning pipeline. The approach is evaluated on the balanced Arabic text dataset provided by the shared task organizers. Our system achieved an F1-score of 0.864 and ranked first among all participating teams.

1 Introduction

The rapid advancement of generative artificial intelligence has significantly transformed the landscape of content creation, education, and communication. State-of-the-art large language models (LLMs) such as GPT-3 (Brown et al., 2020), GPT-4 (OpenAI et al., 2023), and LLaMA (Touvron et al., 2023) are now capable of producing text that exhibits a high degree of fluency, coherence, and stylistic refinement, often closely resembling human writing. These technologies offer substantial benefits, including personalized learning, writing support, and scalable content generation. However, they also raise serious ethical concerns regarding authorship, originality, and academic integrity. Moreover, generative AI can be misused to produce misleading or deceptive content, including fabricated news articles (Ishraquzzaman et al., 2025), deepfake tweets (Fagni et al., 2021), and AI-generated documents such as academic papers and study reports (Chowdhury et al., 2025). Such misuse poses significant ethical risks across domains, including journalism, education, and social media. In light of these developments, there is a growing need and a corresponding challenge to reliably distinguish between human-written and machine-generated text.

Arabic, one of the six official languages recognized worldwide (Wahdan et al., 2020) and the fourth most used language on the Internet with over 400 million speakers (Guellil et al., 2021), has received comparatively less attention in the area of AI-generated text detection. In this context, the AraGenEval shared task (Arabic Authorship Style Transfer and AI-Generated Text Detection) (Abudalfa et al., 2025) is introduced to foster research on Arabic text generation and detection. One of its subtasks, ARATECT, focuses on the binary classification of Arabic texts as either human-written or AI-generated.

To address this challenge, we propose an ensemble-based classification framework that combines the strengths of both multilingual and Arabic-specific pre-trained language models. By integrating Fanar, AraBERT, and XLM-R within a fine-tuning pipeline and applying a majority voting strategy, this approach enhanced the robustness and accuracy of our system, enabling it to rank first among the 16 submitted systems in the ARATECT subtask.

The implementation is publicly available[3] to support transparency and reproducibility.

*Corresponding author: zita.kaoutar@univ-ghardaia.edu.dz

[1] Laboratoire des Mathématiques et Sciences Appliquées, University of Ghardaia, Algeria

[2] https://ezzini.github.io/AraGenEval/

[3] https://github.com/kaoutarzi/AraGenEval-2025-Aratect

Proceedings of The Third Arabic Natural Language Processing Conference, pages 26–31
November 8-9, 2025 ©2025 Association for Computational Linguistics

2 Background

2.1 Task Setup

In this study, we address the detection of AI-generated Arabic text as part of the ARATECT subtask in the AraGenEval Shared Task. This subtask is formulated as a binary classification problem in which the system is given an Arabic text and must determine whether it was written by a human or generated by an AI model. The dataset used consists of Arabic texts spanning various genres, including news articles and literary content. It is balanced in terms of class distribution, featuring an equal number of human- and machine-generated samples. The full dataset comprises 5,798 texts, split into training, development, and test sets, as detailed in Table 1.

For instance, a system might encounter a news excerpt such as:

" قالت وكالة الأنباء السورية سانا إن الدفاعات الجوية السورية تصدت لعدوان إسرائيلي بعدد من الصواريخ استهدفت مناطق في محيط العاصمة دمشق في الساعات الأولى من اليوم الخميس. "

Which means "The Syrian Arab News Agency (SANA) reported that Syrian air defenses responded to an Israeli attack involving several missiles that targeted areas around the capital, Damascus, in the early hours of Thursday." The system is then expected to classify the text accordingly.

2.2 Related Work

Numerous studies (Liu et al., 2025; Wu et al., 2025; Fraser et al., 2025) have addressed the challenge of detecting AI-generated text, driven by the growing capabilities of large language models. However, most existing research has focused predominantly on English or other high-resource languages.

For instance, Katib et al. (2023) introduced a hybrid model called TSA-LSTMRNN, which integrates LSTM with an attention mechanism and the Tunicate Swarm Algorithm (Kaur et al., 2020). They utilize TF-IDF, count vectorizer, and word embeddings for feature extraction, achieving up to 93.83% accuracy in distinguishing between human- and ChatGPT-generated text.

Antoun et al. (2023) proposed a methodology for detecting ChatGPT-generated French text by translating the HC3 English dataset (Guo et al., 2023) and training classifiers (e.g., CamemBERTa, XLM-R). The detectors performed well in-domain (F1 $\approx$ 0.97), but showed reduced effectiveness on out-of-domain and adversarial samples, highlighting limitations in generalization.

Focusing specifically on Arabic, Alshammari et al. (2024) propose two fine-tuned Transformer-based models, AraELECTRA and XLM-R, for detecting AI-generated versus human-written texts. Their approach incorporates a novel Dediacritization Layer. Trained on the AIRABIC dataset (Alshammari and El-Sayed, 2023), the models achieve up to 83% accuracy, outperforming GPTZero (63%) and OpenAI Text Classifier (50%).

Similarly, Alghamdi and Alowibdi (2024) compiled a dataset of Arabic tweets authored by both humans and ChatGPT. They trained and evaluated three machine learning models (SVM, Naive Bayes, and Decision Tree), with Naive Bayes achieving the highest accuracy of 93% in distinguishing between the two sources.

3 System Overview

In this study, we progressively explored a wide range of models for Arabic text classification to address the task of detecting AI-generated content. We began with traditional machine learning methods, advanced through deep learning architectures, and further extended our investigation by fine-tuning various pre-trained language models. To enhance overall performance and robustness, we adopt an ensemble strategy based on majority voting (Dong et al., 2020). The following sections provide a detailed exploration of each category of models employed in our study.

3.1 Machine Learning-based Classification

To classify Arabic AI-generated text using traditional machine learning, we extracted three types of features: (1) statistical and stylistic features, such as word counts, lexical diversity, and punctuation usage; (2) TF-IDF features, which captured sparse lexical patterns; and (3) contextual representations derived from AraBERT embeddings. These features were then used as input to machine learning models, specifically Logistic Regression and a Multi-Layer Perceptron (MLP), which were selected based on their performance on the development set.

3.2 Deep Learning-based Classification

To explore deep learning-based detection, we designed a fusion architecture that integrates both handcrafted and contextual features. As shown in Figure 1, the input text is processed twice to get a rich encoding. The first branch encodes handcrafted stylometric and sparse lexical patterns (stylistic features and TF-IDF), while the second processes semantic features obtained via AraBERT embeddings. This separation aims to preserve the distinct contribution of each feature type and prevent potential dominance of contextual embeddings. The outputs from both branches are then concatenated and passed through a multi-head attention layer to model cross-feature interactions, enabling the integration of both surface-level and deep contextual cues for the final classification.

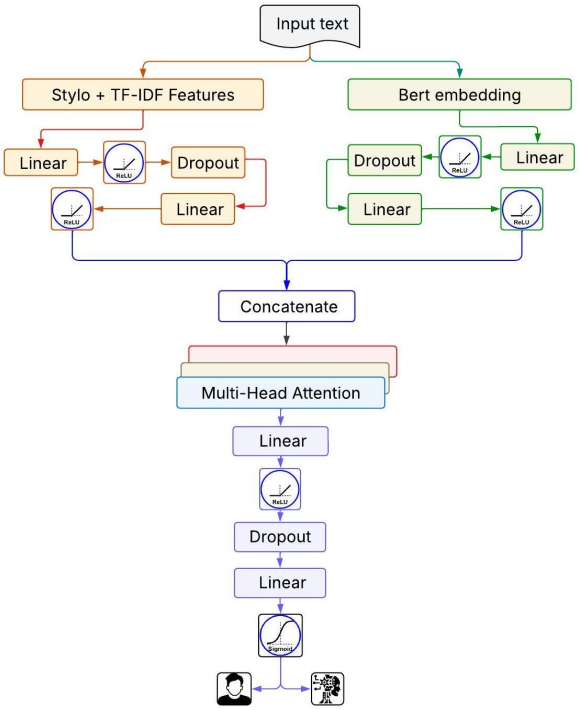

Figure 1: FusionNet Architecture for Arabic AI Generated Text Detection.

3.3 LLM-based Classification

A core focus of our work lies in exploring the potential of large pre-trained language models (LLMs) for detecting AI-generated text. To this end, we experimented with several models and identified three that contributed the most significantly to our final submission results: Fanar, AraBERT, and XLM-R.

XLM-RoBERTa[4] is a multilingual transformer-based language model developed to handle over 100 languages, including Arabic. It builds upon the RoBERTa architecture and is trained using the Masked Language Modeling (MLM) objective on a massive dataset of 2.5TB of filtered CommonCrawl data. Its architecture supports fine-tuning for tasks such as text classification, sentiment analysis, and question answering, leveraging rich contextual representations learned from diverse multilingual corpora (Conneau et al., 2020).

AraBERT[5] is a transformer-based language model specifically pre-trained for Arabic, adapting the original BERT (Devlin et al., 2019) architecture to better address the linguistic richness and morphological complexity of Arabic. Trained on approximately 1.5 billion words from diverse Arabic corpora, AraBERT demonstrates strong performance across various NLP tasks such as sentiment analysis, question answering, and named entity recognition. Its design, which includes 12 encoder layers and 136M parameters, allows it to capture deep contextual representations tailored to the Arabic language (Antoun et al., 2020).

Fanar[6] is an Arabic-centric multimodal Large Language Model developed by the Qatar Computing Research Institute at Hamad Bin Khalifa University. It is available in two versions: Fanar Star (7B) and Fanar Prime (9B), trained on a corpus of one trillion tokens in Arabic and English. Fanar is designed to support Modern Standard Arabic as well as major regional dialects. Aligned with Islamic values and Arab cultural contexts, it offers a range of capabilities such as text generation, speech and image processing, and retrieval-augmented generation (RAG) (Team et al., 2025).

Finally, as shown in Figure 2, the predictions from the fine-tuned XLM-RoBERTa, AraBERT, and Fanar models were combined using a majority voting scheme. This ensemble method leveraged the complementary strengths of the individual models to achieve balanced performance across all evaluation metrics and improve the overall accuracy and robustness of the text classification system.

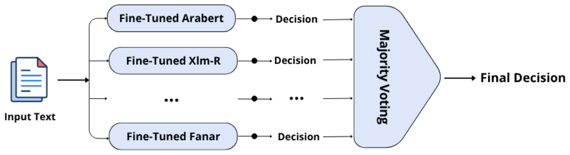

Figure 2: Ensemble-Based Approach for Arabic AI-Generated Text Detection.

4 Experimental Setup

We deploy the dataset provided in the ARATECT subtask of the AraGenEval shared task (Abudalfa

[4] FacebookAI/xlm-roberta-base

[5] aubmindlab/bert-base-arabertv2

[6] https://huggingface.co/QCRI/Fanar-1-9B

et al., 2025), which aims to detect AI-generated Arabic texts. The dataset comprises a balanced set of human- and machine-generated texts across the training, development, and test splits. Human-written texts were sourced from credible Arabic news platforms and literary works, ensuring diversity in style and topic. In contrast, machine-generated texts were produced using multiple large language models, including Mistral, GPT-4, and LLaMA.

Table 1 provides a detailed overview of the dataset's composition.

Data	Training	Dev	Test
# of Samples	4,798	500	500
# of Words	2,330,765	139,745	115,057
Machine (%)	50%	50%	50%
Human (%)	50%	50%	50%

Table 1: ARATECT Dataset Overview.

All experiments were conducted using Python within a Kaggle GPU environment, leveraging the Hugging Face Transformers, Datasets, and Evaluate libraries to fine-tune three pre-trained language models: XLM-RoBERTa, AraBERT, and Fanar. For XLM-RoBERTa and AraBERT, texts were tokenized and classified using cross-entropy loss, with a batch size of 4 over 3 epochs and 1 epoch, respectively. Fanar was fine-tuned using instruction-formatted prompts through LoRA-based parameter-efficient tuning in 4-bit precision, with a batch size of 2 and one epoch. Model performance was evaluated using accuracy, precision, recall, and F1-score. All implementation details, including code and configurations, are publicly available on GitHub[7].

5 Results

Table 2 presents the evaluation results across all experimented models. Traditional machine learning approaches and FusionNet obtained relatively modest performance, reflecting their limited ability to capture the complex linguistic patterns in the dataset. Among the Transformer-based models, the three fine-tuned large language models XLM-R, AraBERT, and Fanar stood out with superior and complementary strengths. AraBERT achieved the highest accuracy (0.864) and F1-score (0.861), XLM-R attained the highest precision (0.911), and Fanar recorded the highest recall (0.920). Although the performance of the Majority Voting ensemble is numerically close to that of AraBERT, the ensemble remains valuable because it balances these strengths, producing a more stable and robust system that is less dependent on the behavior of a single model and better suited to varying data distributions.

Model	Acc.	Prec.	Rec.	F1
LR	0.438	0.464	0.804	0.589
MLP	0.506	0.503	0.988	0.667
FusionNet	0.578	0.552	0.824	0.661
AraElectra	0.688	0.737	0.584	0.652
MARBERT	0.586	0.563	0.764	0.649
DeBERTa	0.768	0.791	0.728	0.758
Qwen2.5	0.480	0.490	0.940	0.644
CAMeL	0.642	0.612	0.776	0.684
XLM-R	0.832	0.911	0.736	0.814
AraBERT	0.864	0.882	0.840	0.861
Fanar	0.776	0.714	0.920	0.804
Majority Voting	**0.866**	**0.877**	**0.852**	**0.864**

Table 2: Performance of our models.

6 Conclusion

In this study, we developed a system for AI-generated Arabic text detection within the ARATECT subtask of the AraGenEval Shared Task. We proposed an ensemble-based classification framework that combines the strengths of both multilingual and Arabic-specific pre-trained language models. By integrating Fanar, AraBERT, and XLM-R within a fine-tuning pipeline and applying a majority voting strategy, the system achieved strong and balanced performance across all evaluation metrics. However, there is room for improvement, particularly in enhancing generalization capabilities to unseen domains and handling more diverse writing styles. Future work will address these limitations by exploring more advanced ensemble learning techniques, such as stacking, incorporating larger and more recent language models like GPT-4 or LLaMA 3, and evaluating the system on broader datasets to further improve robustness and adaptability. Furthermore, we plan to extend the classification task beyond binary detection to detect specific AI-generated segments within texts.

[7] https://github.com/kaoutarzi/AraGenEval-2025-Aratect

Acknowledgments

We gratefully acknowledge the organizers of the AraGenEval Shared Task for providing the dataset and evaluation framework used in this study. This work is supported by the Algerian Directorate-General for Scientific Research and Technological Development (DGRSDT), and performed under the PRFU Projects: C00L07N030120220002 and C00L07UN470120230001.

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.

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.

Hamed Alshammari and Ahmed El-Sayed. 2023. AIRABIC: Arabic Dataset for Performance Evaluation of AI Detectors. In *2023 International Conference on Machine Learning and Applications (ICMLA)*, pages 864–870.

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).

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, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

Wissam Antoun, Virginie Mouilleron, Benoît Sagot, and Djamé Seddah. 2023. Towards a robust detection of language model-generated text: Is ChatGPT that easy to detect? In *Actes de CORIA-TALN 2023. Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles (TALN), volume 1 : travaux de recherche originaux – articles longs*, pages 14–27, Paris, France. ATALA.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Shammur Absar Chowdhury, Hind Almerekhi, Mucahid Kutlu, Kaan Efe Keleş, Fatema Ahmad, Tasnim Mohiuddin, George Mikros, and Firoj Alam. 2025. GenAI content detection task 2: AI vs. human – academic essay authenticity challenge. In *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)*, pages 323–333, Abu Dhabi, UAE. International Conference on Computational Linguistics.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

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, Minneapolis, Minnesota. Association for Computational Linguistics.

Xibin Dong, Zhiwen Yu, Wenming Cao, Yifan Shi, and Qianli Ma. 2020. A survey on ensemble learning. *Front. Comput. Sci.*, 14(2):241–258.

Tiziano Fagni, Fabrizio Falchi, Margherita Gambini, Antonio Martella, and Maurizio Tesconi. 2021. Tweepfake: About detecting deepfake tweets. *PLOS ONE*, 16(5):1–16.

Kathleen Fraser, Hillary Dawkins, and Svetlana Kiritchenko. 2025. Detecting AI-Generated Text: Factors Influencing Detectability with Current Methods. *Journal of Artificial Intelligence Research*, 82:2233–2278.

Imane Guellil, Houda Saâdane, Faical Azouaou, Billel Gueni, and Damien Nouvel. 2021. Arabic natural language processing: An overview. *Journal of King Saud University - Computer and Information Sciences*, 33(5):497–507.

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How Close is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection. *ArXiv*, abs/2301.07597.

Md Ishraquzzaman, Ashraful Islam, Shahreen Rahman, and Riasat Khan. 2025. Ensemble Transformer-Based Detection of Fake and AI-Generated News.

Applied Computational Intelligence and Soft Computing, 2025.

Iyad Katib, Fatmah Y. Assiri, Hesham A. Abdushkour, Diaa Hamed, and Mahmoud Ragab. 2023. Differentiating Chat Generative Pretrained Transformer from Humans: Detecting ChatGPT-Generated Text and Human Text Using Machine Learning. *Mathematics*, 11(15).

Satnam Kaur, Lalit K. Awasthi, A.L. Sangal, and Gaurav Dhiman. 2020. Tunicate Swarm Algorithm: A new bio-inspired based metaheuristic paradigm for global optimization. *Engineering Applications of Artificial Intelligence*, 90:103541.

Xin Liu, Yang Li, and Kan Li. 2025. Enhancing the Robustness of AI-Generated Text Detectors: A Survey. *Mathematics*, 13(13).

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and Barret Zoph. 2023. GPT-4 Technical Report.

Fanar Team, Ummar Abbas, Mohammad Shahmeer Ahmad, Firoj Alam, Enes Altinisik, Ehsannedin Asgari, Yazan Boshmaf, Sabri Boughorbel, Sanjay Chawla, Shammur Chowdhury, Fahim Dalvi, Kareem Darwish, Nadir Durrani, Mohamed Elfeky, Ahmed Elmagarmid, Mohamed Eltabakh, Masoomali Fatehkia, Anastasios Fragkopoulos, Maram Hasanain, and 23 others. 2025. Fanar: An Arabic-Centric Multimodal Generative AI Platform. *Preprint*, arXiv:2501.13944.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models.

Ahlam Wahdan, Sendeyah Hantoobi, Said Salloum, and Khaled Shaalan. 2020. A systematic review of text classification research based on deep learning models in Arabic language. *International Journal of Electrical and Computer Engineering (IJECE)*, pages 6629–6643.

Junchao Wu, Shu Yang, Runzhe Zhan, Yulin Yuan, Lidia Sam Chao, and Derek Fai Wong. 2025. A survey on LLM-generated text detection: Necessity, methods, and future directions. *Computational Linguistics*, 51(1):275–338.