BUSTED at AbjadGenEval Shared Task: LoRAD: Low-Resource AI-Generated Text Detection with XLM-RoBERTa

Ali Zain
vin.alizain@gmail.com
Karachi, Pakistan

Abstract

This paper describes our system submitted to the AbjadGenEval Shared Task at ArabicNLP 2026, which focuses on binary classification of human-written versus machine-generated text in low-resource languages. We participated in two independent subtasks targeting Arabic and Urdu news and literary texts. Our approach relies exclusively on fine-tuning XLM-RoBERTa, a multilingual Transformer-based model, under carefully controlled training and preprocessing settings. While the same model architecture was used for both subtasks, language-specific data handling strategies were applied based on empirical observations. The proposed system achieved first place in the Urdu subtask and third place in the Arabic subtask according to the official evaluation. These results demonstrate that multilingual pretrained models can serve as strong and reliable systems for AI-generated text detection across diverse languages.

1 Introduction

The increasing use of large language models (LLMs) for automated content generation has created a growing demand for reliable methods to distinguish machine-generated text from human-authored writing. This challenge is particularly acute for low-resource languages, where linguistic complexity and limited annotated data complicate detection efforts.

The AbjadGenEval Shared Task (Ezzini et al., 2026) , organized as part of ArabicNLP 2026, addresses this challenge by providing standardized benchmarks for AI-generated text detection in multiple languages. In this paper, we describe our participation in two independent subtasks: Arabic AI-generated text detection and Urdu AI-generated text detection. Both subtasks are formulated as binary classification problems, but differ in language structure, dataset composition, and empirical behavior.

Rather than proposing new model architectures, our submission focuses on disciplined system design using a single multilingual backbone. We employ XLM-RoBERTa (Conneau et al., 2020) for both subtasks and explore how language-specific preprocessing and data usage choices influence performance within a shared-task setting.

2 Related Work

The problem of distinguishing human-authored text from machine-generated content has long been studied under the broader umbrella of authorship attribution and stylometric analysis. Early approaches relied primarily on manually engineered features such as character and word n-grams, lexical richness measures, readability indices, and syntactic patterns. These feature-based methods were shown to be effective for detecting outputs from early statistical and rule-based generators, but their robustness degrades substantially as modern neural language models produce increasingly fluent and contextually coherent text (Zellers et al., 2019; Ippolito et al., 2020).

With the emergence of large pretrained language models, research has largely shifted toward neural approaches. Fine-tuning pretrained transformers, particularly BERT-style architectures (Liu et al., 2019), has become a strong and widely adopted baseline for machine-generated text detection across domains and languages. These models implicitly capture subtle distributional and stylistic regularities that are difficult to encode manually, making them better suited for detecting high-quality neural text. Multilingual variants such as XLM-RoBERTa (Conneau et al., 2020) have further enabled cross-lingual transfer, which is especially valuable for under-resourced languages where labeled data is limited.

Beyond standard fine-tuning, several studies have explored detecting artifacts that arise from

Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script (AbjadNLP 2026), pages 468–471

the generative process itself. Statistical methods analyze token probability distributions, repetition patterns, and entropy-based signals that differ between human and machine text (Gehrmann et al., 2019). More recent work proposes model-agnostic detection techniques based on curvature in likelihood space, demonstrating that neural generators leave implicit statistical fingerprints even when outputs appear highly natural (Mitchell et al., 2023). In parallel to detection-based approaches, watermarking techniques embed signals during text generation, although such methods require control over the generation process and are therefore less applicable in open-world scenarios.

In the context of Arabic and other morphologically rich languages, AI-generated text detection remains comparatively under-explored. Several shared tasks have recently addressed this gap by providing benchmark datasets and standardized evaluation protocols. The AraGenEva(Abudalfa et al., 2025) shared task focused on Arabic AI-generated text detection and demonstrated that multilingual transformers often outperform Arabic-specific models, highlighting the benefit of cross-lingual pretraining (Zain et al., 2025). Similarly, the M-DAIGT shared task (Lamsiyah et al., 2025) expanded the scope to multi-domain detection, including news and academic writing, and showed that transformer-based systems consistently outperform classical pipelines across domains (Farooqui et al., 2025).

Our work is situated within this recent line of shared-task-driven research. We adopt the fine-tuning paradigm using XLM-RoBERTa and extend prior findings to the AbjadGenEval shared task, which focuses on Arabic and Urdu. Unlike approaches that rely heavily on aggressive text normalization, we explore how preserving raw textual characteristics can aid detection, particularly in low-resource settings. By participating in both language-specific subtasks under a unified framework, our system contributes further empirical evidence on the effectiveness of multilingual transformers for AI-generated text detection in under-represented languages.

3 Task Description

Our work addresses the AbjadGenEval Shared Task on AI-generated text detection, organized as part of the ArabicNLP 2026 workshop. The primary objective of this shared task is to develop systems capable of distinguishing between human-written and machine-generated texts for languages that use the Abjad script. The shared task provides standardized datasets and evaluation protocols on Codabench, and invites participants to compete on separate language tracks. The official task description and resources are available online at the task website https://ezzini.github.io/AbjadGenEval/.

The shared task consists of two main binary classification subtasks corresponding to individual languages:

- **Arabic Subtask:** Detect whether an Arabic text is human-written or machine-generated. Evaluation for this subtask was hosted on Codabench [1].
- **Urdu Subtask:** Detect whether an Urdu text is human-written or machine-generated. Evaluation for this subtask was hosted on Codabench [2].

Each subtask is evaluated independently using official datasets and metrics provided by the organizers. The primary evaluation metric is macro F1-score, with accuracy reported as a secondary indicator of performance. Although the general objective (binary classification) is consistent across subtasks, they are treated as separate experiments due to differences in script, language morphology, and data distributions.

4 Data and Preprocessing

4.1 Arabic Subtask

For the Arabic subtask, we used the official dataset released by the task organizers and augmented it with an external dataset of AI-generated Arabic abstracts, also provided by the organizers and hosted on Hugging Face.[3]

The resulting training data combines in-domain labeled examples with additional machine-generated and human-written samples to improve robustness against stylistic variation. The composition of the Arabic datasets is summarized in Table 1.

All samples were assigned binary labels using the convention 0 for human-written text and 1 for machine-generated text. We removed missing or

[1] https://www.codabench.org/competitions/12306/
[2] https://www.codabench.org/competitions/12319/
[3] https://huggingface.co/datasets/KFUPM-JRCAI/arabic-generated-abstracts

empty entries during preprocessing. Tokenization was performed using the XLM-RoBERTa tokenizer with a maximum sequence length of 256 tokens. Padding was applied dynamically at the batch level to ensure computational efficiency.

4.2 Urdu Subtask

For the Urdu subtask, we strictly relied on the dataset released by the task organizers and did not incorporate any external data sources. The dataset consists of 11,904 training samples and 2,630 test samples, with balanced class distributions.

In contrast to conventional preprocessing pipelines, we deliberately avoided aggressive text normalization or cleaning. This decision was motivated by the hypothesis that subtle surface-level patterns and distributional irregularities introduced by neural generators may serve as implicit signals for detection. Preserving the original text allows the model to exploit such cues during fine-tuning. Tokenization was performed with a maximum sequence length of 512 tokens in order to accommodate longer document structures commonly observed in Urdu news and literary text.

Subtask	Split	Human	Machine	Total
Arabic	Train	2,639	2,639	5,278
Arabic	External Data	2,574	10,296	12,870
Arabic	Test	–	–	200
Urdu	Train	5,952	5,952	11,904
Urdu	Test	–	–	2,630

Table 1: Dataset statistics for the Arabic and Urdu subtasks. The Arabic external dataset was used only for training.

5 System Overview

Our system is based on XLM-RoBERTa (Conneau et al., 2020), a multilingual Transformer pretrained on large-scale CommonCrawl data covering more than 100 languages. A standard classification head was added on top of the [CLS] representation for binary prediction.

The same model architecture was used for both subtasks. No task-specific architectural changes or auxiliary objectives were introduced, allowing us to isolate the impact of data handling and training configuration.

6 Experimental Setup

All experiments were conducted in a Kaggle environment using dual NVIDIA T4 GPUs. Due to GPU memory constraints, batch sizes varied across runs.Table 2.

Hyperparameter	Value
Learning Rate	2e-5
Batch Size	4–16 (depending on GPU memory)
Optimizer	AdamW
Weight Decay	0.01
Epochs	Upto 6

Table 2: Key hyperparameters for fine-tuning.

Each subtask was trained independently using its respective dataset and tokenization settings.

7 Results

Evaluation was performed using the official shared-task evaluation scripts. The primary metric is F1-score, with Accuracy reported as a secondary metric. Precision and Recall are also included for completeness.

7.1 Urdu Subtask (Ranked 1st)

Metric	Score
F1-score	0.8877
Accuracy	0.8878
Precision	0.8834
Recall	0.8922

7.2 Arabic Subtask (Ranked 3rd)

Metric	Score
F1-score	0.8866
Accuracy	0.8850
Precision	0.8737
Recall	0.9000

8 Discussion

Despite relying on a single multilingual model, the system achieved strong performance across both subtasks. The results indicate that XLM-RoBERTa can effectively adapt to different languages when combined with appropriate data handling strategies.

Empirical observations suggest that preprocessing decisions play an important role. In particular, preserving raw text characteristics proved beneficial for the Urdu subtask, while data augmentation contributed to robustness in the Arabic subtask. These findings highlight the importance of language-aware pipeline design within shared-task environments.

Limitations

This work focuses exclusively on a single multilingual model and does not include comparisons with alternative monolingual or multilingual baselines. Additionally, evaluation is limited to the datasets provided by the shared task and one external Arabic dataset. Future work should explore cross-domain generalization and robustness to newer generation models.

Acknowledgments

We thank the organizers of the AbjadGenEval Shared Task for providing the datasets and evaluation framework.

Code availability: *[Placeholder: to be released].*

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.

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.

Saad Ezzini, Irfan Ahmed, Salmane Chafik, Shadi Abudalfa, Mo El-Haj, Ahmed Abdelali, Mustafa Jarrar, Nadir Durrani, Hassan Sajjad, and Farah Adeeba. 2026. Abjadgeneval: Abjad ai generated text detection shared task for languages using arabic script 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.

Sareem Farooqui, Ali Zain, and Muhammad Rafi. 2025. Shared task on multi-domain detection of ai-generated text (m-daigt). In *Proceedings of the Shared Task on Multi-Domain Detection of AI-Generated Text*, pages 15–19. Accessed January 2026.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M. Rush. 2019. Gltr: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 111–116.

Daphne Ippolito, Daniel Duckworth, Chris Callison-Burch, and Douglas Eck. 2020. Automatic detection of generated text is easiest when humans are fooled. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1808–1822.

Salima Lamsiyah, Saad Ezzini, Abdelkader El Mahdaouy, Hamza Alami, Abdessamad Benlahbib, Samir El amrany, Salmane Chafik, and Hicham Hammouchi. 2025. M-DAIGT: A Shared Task on Multi-Domain Detection of AI-Generated Text. In *Proceedings of the Shared Task on Multi-Domain Detection of AI-Generated Text*, pages 1–9, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D. Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. *arXiv preprint arXiv:2301.11305*.

Ali Zain, Sareem Farooqui, and Muhammad Rafi. 2025. Busted at arageneval shared task: A comparative study of transformer-based models for arabic ai-generated text detection. In *Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks*, pages 72–76. Accessed January 2026.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. In *Advances in Neural Information Processing Systems*.