XLMR-Urdu at AbjadGenEval Shared Task: A Data-Centric Transformer-Based Approach for AI-Generated Urdu Text Detection

Mohannad Hendi
Independent Researcher
M.hendi1@student.aaup.edu

Abstract

The rapid advancement of large language models (LLMs) has led to a substantial increase in automatically generated textual content, raising concerns regarding misinformation, plagiarism, and authorship verification. These challenges are particularly pronounced for low-resource languages such as Urdu, where limited annotated data and complex linguistic properties hinder robust detection. In this paper, we present a transformer-based approach for binary classification of human-written versus AI-generated Urdu text, developed for the AbjadGenEval Task 2 shared task. Beyond model fine-tuning, we adopt a data-centric perspective, emphasizing dataset diagnostics, document-level inference, and calibration strategies. Our system achieves strong performance on the official test set, with an F1-score of 88.68% and balanced accuracy of 88.71%. Through empirical analysis, we demonstrate that dataset characteristics and generator-specific artifacts play a dominant role in model generalization, highlighting critical directions for future research in low-resource AI-generated text detection.

1 Introduction

The rapid proliferation of large language models (LLMs) such as GPT-3 (Brown et al., 2020), GPT-4 (Achiam et al., 2023), and open-source alternatives like LLaMA (Touvron et al., 2023) has fundamentally transformed text production across diverse domains including journalism, education, creative writing, and online platforms. While these models enable unprecedented efficiency in content generation, they simultaneously introduce critical challenges related to misinformation (Zellers et al., 2019), academic plagiarism (Liang et al., 2023), and the broader erosion of trust in written content (Kreps et al., 2022). Consequently, automatic detection of AI-generated text has emerged as a crucial research problem, typically formulated as a binary classification task distinguishing human-written from machine-generated content.

Recent shared-task initiatives have played a central role in benchmarking progress on this problem, particularly for languages using Arabic script, by providing standardized datasets, evaluation protocols, and comparative baselines (Ezzini et al., 2026).

Recent studies have explored diverse detection approaches, ranging from zero-shot statistical methods based on perplexity and entropy (Mitchell et al., 2023) to supervised neural classifiers fine-tuned on labeled datasets (Solaiman et al., 2019; Gehrmann et al., 2019). Comprehensive surveys indicate that transformer-based detectors often achieve strong in-domain performance when trained on sufficient annotated data (Wu et al., 2025; Kumarage et al., 2024; Ippolito et al., 2020). However, multiple investigations reveal fundamental brittleness under distribution shifts, paraphrasing attacks (Krishna et al., 2023), and exposure to unseen generation strategies (Sadasivan et al., 2023), resulting in severely degraded generalization to real-world deployment scenarios.

The vast majority of existing research concentrates on high-resource languages, particularly English, where large-scale annotated datasets (Wang et al., 2024) and extensive model resources are readily available. In stark contrast, low-resource languages remain critically underexplored despite facing equal or greater risks from AI-generated misinformation and content manipulation. Urdu, an Indo-Aryan language spoken by over 230 million people worldwide, presents distinctive challenges including its Perso-Arabic cursive script, complex morphological structure, and severe scarcity of curated NLP datasets (Ali et al., 2008). To date, only a limited number of studies explicitly examine AI-generated text detection

Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script (AbjadNLP 2026), pages 461–467

in Urdu or closely related South Asian languages, typically leveraging multilingual pretrained models with varying degrees of success (Ammar et al., 2025). Related shared tasks such as AraGenEval (Abudalfa et al., 2025) further highlight the challenges of AI-generated text detection in Arabic and related languages, emphasizing the role of generator diversity and evaluation design in shaping reported performance.

Multilingual transformer architectures such as XLM-RoBERTa (Conneau et al., 2020) and mBERT (Devlin et al., 2019) provide competitive baselines for cross-lingual text classification tasks, including human versus AI discrimination across multiple languages (Ali et al., 2025; Schaaff et al., 2023). Nevertheless, accumulating evidence suggests that reported detection performance may be substantially inflated by dataset artifacts (Gehrmann et al., 2019), generator-specific lexical and stylistic fingerprints (Ippolito et al., 2020), and evaluation protocols that inadequately reflect authentic deployment conditions. These concerns are further corroborated by multi-domain evaluation efforts such as the M-DAIGT shared task (Lamsiyah et al., 2025), which demonstrate that domain shifts and dataset composition critically impact detector robustness.

In this work, we address AI-generated Urdu text detection within the framework of the **AbjadGenEval Task 2** shared task, which focuses on document-level classification of Urdu news articles. Rather than proposing complex architectural innovations, we deliberately adopt a **data-centric perspective** that prioritizes careful dataset analysis, robust document-level inference strategies, systematic handling of long-form text, and probability calibration techniques to mitigate class-specific misclassification bias.

Our primary contributions are threefold:

- We conduct a comprehensive data diagnostic analysis of the AbjadGenEval Task 2 dataset, revealing distributional characteristics, lexical patterns, and document length variations that inform our system design choices and highlight the presence of dataset-specific artifacts.

- We develop a detection system based on XLM-RoBERTa with sliding window segmentation (L=256, S=64), document-level aggregation via mean pooling, differential learning rates, and calibrated threshold optimization, achieving an F1-score of 88.68% and balanced accuracy of 88.71% on the official test set. Through error analysis, we demonstrate that dataset composition and generator-specific patterns significantly influence detection performance in low-resource language settings.

2 Related Work

Research on AI-generated text detection has expanded rapidly with the widespread adoption of large language models. Early approaches relied on statistical and stylometric features, such as lexical diversity, sentence length distributions, and perplexity-based measures. While effective in constrained settings, these methods generally lack robustness and struggle to generalize across domains and generation styles.

Recent work has largely shifted toward neural approaches, particularly transformer-based classifiers fine-tuned for binary detection tasks. Surveys of AI-generated text detection systems report that pretrained language models achieve strong in-domain performance when sufficient labeled data is available (Wu et al., 2025; Kumarage et al., 2024). However, several studies have demonstrated that such detectors often exploit superficial patterns and can fail under distribution shifts, paraphrasing, or adversarial rewriting (Sadasivan et al., 2023).

Most existing detection research focuses on high-resource languages, especially English. In contrast, low-resource languages remain underrepresented despite facing similar risks from AI-generated misinformation. Only a limited number of studies explicitly address AI-generated text detection for Urdu. Ammar et al. (Ammar et al., 2025) investigate detection in Urdu using multilingual pretrained models and report promising results, while also highlighting sensitivity to dataset composition and generator diversity.

Multilingual transformer models, such as XLM-RoBERTa, have been evaluated for AI-generated text detection across multiple languages. Prior work shows that these models provide strong cross-lingual representations and competitive baselines for human versus AI discrimination (Ali et al., 2025; Schaaff et al., 2023). Nevertheless, these studies also suggest that performance may be influenced by dataset artifacts and

generator-specific cues, raising concerns about real-world generalization.

In contrast to prior work that emphasizes architectural modifications or model complexity, our approach adopts a data-centric perspective. We focus on document-level inference, robust handling of long texts, and calibration strategies to mitigate misclassification bias. By combining transformer-based modeling with detailed data diagnostics, we aim to better characterize the limitations of current detection approaches for low-resource languages such as Urdu.

3 Dataset and Data Diagnostics

3.1 Dataset Description

The experiments in this work are conducted using the dataset released as part of the **AbjadGenEval Task 2** shared task, which focuses on distinguishing *human-written* and *AI-generated* Urdu news text. The task is formulated as a *binary document-level classification problem*, where each instance corresponds to a complete news article rather than isolated sentences.

The training set consists of **11,910 documents**, each annotated with a binary label, where label 0 denotes human-written text and label 1 denotes AI-generated text. All training samples are long-form Urdu news articles. In addition to the raw textual content, the dataset provides a pre-computed `word_count` attribute representing the number of whitespace-separated tokens per document, enabling direct analysis of document length characteristics.

The official test set contains **2,630 unlabeled documents** and is used exclusively for final evaluation. Only the textual content is provided for test instances, without labels or auxiliary metadata.

3.2 Label Distribution

Inspection of the training data indicates that the dataset is **approximately balanced** across the two classes, with comparable numbers of human-written and AI-generated documents. This class balance reduces the risk of biased learning toward a dominant class and allows evaluation metrics such as accuracy, balanced accuracy, and F1-score to be interpreted reliably without requiring aggressive resampling or class-weighting strategies, as also noted in prior shared-task evaluations (Wu et al., 2025).

3.3 Text Length Characteristics

Analysis of document length reveals that the dataset primarily consists of **long-form news articles**. Most documents contain more than **300 tokens**, with a substantial portion extending well beyond this threshold. This property differentiates the task from sentence-level AI-generated text detection and introduces challenges related to the fixed input length constraints of transformer-based models.

Figure 1 illustrates the distribution of legacy token lengths for human-written and AI-generated documents. While the two distributions exhibit significant overlap, AI-generated documents tend to display a **more concentrated length distribution**, whereas human-written articles show **greater variance** in token counts. Similar observations have been reported in prior analyses of AI-generated text detection datasets (Sadasivan et al., 2023).

These findings suggest that document length alone is insufficient for reliable classification, but may act as a weak auxiliary signal that detection models can implicitly exploit when combined with lexical and stylistic features.

3.4 Lexical and Structural Patterns

To further investigate lexical differences between classes, we analyze token frequency patterns and visualize the most frequent terms using class-specific word clouds, shown in Figure 2. The visualizations reveal partial overlap in topical vocabulary, consistent with both classes covering similar news domains.

However, AI-generated text exhibits **more repetitive usage of high-frequency tokens** and stylistically uniform phrasing, while human-written text demonstrates **greater lexical diversity**, including informal expressions, quotations, and variable narrative flow. In addition, certain tokens and short n-grams appear disproportionately in AI-generated samples, indicating the presence of **generator-specific artifacts**. These patterns align with findings from prior work on multilingual and low-resource AI-generated text detection (Ammar et al., 2025; Ali et al., 2025).

3.5 Data Diagnostics and Implications

Overall, the dataset exhibits **partially overlapping but non-identical distributions** between human-written and AI-generated text. While this

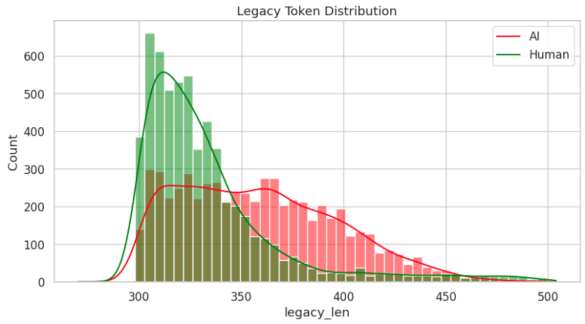

Figure 1: Legacy token length distribution for human-written and AI-generated Urdu documents.

Figure 2: Word cloud visualization for human-written (left) and AI-generated (right) Urdu text.

enables effective in-domain learning, it also introduces the risk that models may overfit to dataset-specific or generator-specific artifacts rather than learning generalizable properties of AI-generated language, a limitation highlighted in recent robustness studies (Sadasivan et al., 2023; Kumarage et al., 2024).

Furthermore, the document-level nature of the data necessitates careful handling of long sequences. Direct truncation would discard substantial contextual information, motivating the use of **sliding window segmentation** and **document-level aggregation** in our modeling approach, which has been adopted in similar document-level detection settings (Ali et al., 2025).

These data diagnostics underscore the critical role of dataset characteristics in determining detection performance and reinforce the importance of **data-centric evaluation and preprocessing strategies** when studying AI-generated text detection in low-resource languages such as Urdu.

4 Methodology

4.1 Model Architecture

Our system is based on fine-tuning a pretrained multilingual transformer for binary classification. We employ XLM-RoBERTa-base as the backbone due to its strong cross-lingual representations and native support for Urdu.

Given a tokenized input sequence $x = (x_1, \ldots, x_n)$, the encoder produces contextualized representations:

$$H = \text{XLM-R}(x), \quad H \in \mathbb{R}^{n \times 768}$$

The pooled representation corresponding to the classification token is passed to a linear classification head:

$$\hat{y} = \text{softmax}(W h_{\text{CLS}} + b)$$

4.2 Long-Document Handling

Since many documents exceed the transformer token limit, we adopt a sliding window segmentation strategy. Each document is divided into overlapping segments of length $L = 256$ tokens with stride $S = 64$. All segments inherit the original document label.

To prevent data leakage, document-level splitting into training and validation sets is performed prior to segmentation, ensuring that no segments from the same document appear in both splits.

4.3 Training Strategy

The model is trained using cross-entropy loss and the AdamW optimizer. To stabilize fine-tuning on limited data, the encoder is frozen during the first epoch and unfrozen thereafter. Differential learning rates are applied, with a lower rate for the encoder and a higher rate for the classification head. Early stopping based on validation F1-score is used to mitigate overfitting.

4.4 Document-Level Aggregation and Calibration

During inference, predictions are produced at the segment level. Given segment probabilities $\{p_1, \ldots, p_k\}$ for a document, the document-level probability is computed as:

$$P(d) = \frac{1}{k} \sum_{i=1}^{k} p_i$$

Instead of using a fixed threshold, we perform threshold optimization on the validation set to maximize F1-score. The final prediction is obtained by comparing $P(d)$ against the optimized threshold.

5 Experiments and Results

We evaluate our model on the official test set using standard binary classification metrics. The results are shown in Table 1.

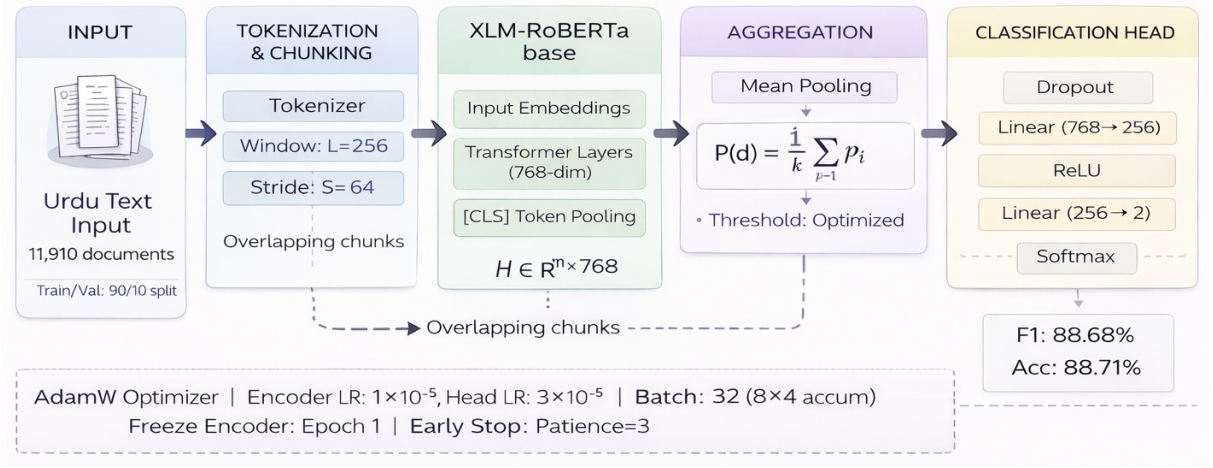

Figure 3: **Model Architecture and Processing Pipeline.** Overview of our Urdu AI-generated text detection system. Documents are segmented using a sliding window (L=256, S=64), encoded with XLM-RoBERTa-base, aggregated at the document level via mean pooling, and classified into human-written or AI-generated text.

Metric	Score (%)
Accuracy	88.71
Balanced Accuracy	88.71
Precision	88.44
Recall	88.91
F1-score	88.68

Table 1: Performance on the AbjadGenEval Task 2 test set.

The close alignment between accuracy and balanced accuracy indicates stable performance across classes. The high recall suggests effective identification of AI-generated text.

6 Error Analysis and Discussion

Qualitative error analysis reveals systematic failure cases. Human-written text exhibiting formal or templated language is sometimes misclassified as AI-generated. Conversely, AI-generated text with increased stylistic variability may be labeled as human-written.

These errors reinforce the conclusion that dataset diversity and generator variation play a critical role in detection performance. Architectural improvements alone are unlikely to resolve these limitations without more representative datasets.

7 Conclusion

We presented a transformer-based approach for AI-generated Urdu text detection, emphasizing a data-centric methodology. Our system achieves strong performance in the AbjadGenEval Task 2 shared task, with an F1-score of 88.68%.

More importantly, our analysis highlights that dataset artifacts and generator-specific patterns dominate performance in low-resource detection tasks. Future work will focus on cross-generator evaluation, adversarial data augmentation, and improved dataset construction to enable more robust and generalizable AI-generated text detection systems.

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.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Iqra Ali, Jesse Atuhurra, Hidetaka Kamigaito, and Taro Watanabe. 2025. Hlu: Human vs llm generated text detection dataset for urdu at multiple granularities. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3495–3510.

Mohammad Naveed Ali, Mohammad Abid Khan, and Muhammad Aamir Khan. 2008. An optimal order of factors for the computational treatment of personal

anaphoric devices in urdu discourse. In *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*.

Muhammad Ammar, Hadiya Murad Hadi, and Usman Majeed Butt. 2025. Ai-generated text detection in low-resource languages: A case study on urdu. *arXiv preprint arXiv:2510.16573*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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.

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.

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.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush. 2019. Gltr: Statistical detection and visualization of generated text. *arXiv preprint arXiv:1906.04043*.

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.

Sarah Kreps, R Miles McCain, and Miles Brundage. 2022. All the news that's fit to fabricate: Ai-generated text as a tool of media misinformation. *Journal of experimental political science*, 9(1):104–117.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. 2023. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. *Advances in Neural Information Processing Systems*, 36:27469–27500.

Tharindu Kumarage, Garima Agrawal, Paras Sheth, Raha Moraffah, Aman Chadha, Joshua Garland, and Huan Liu. 2024. A survey of ai-generated text forensic systems: Detection, attribution, and characterization. *arXiv preprint arXiv:2403.01152*.

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.

Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. 2023. Gpt detectors are biased against non-native english writers. *Patterns*, 4(7).

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. 2023. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International conference on machine learning*, pages 24950–24962. PMLR.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2023. Can ai-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*.

Kristina Schaaff, Tim Schlippe, and Lorenz Mindner. 2023. Classification of human-and ai-generated texts for english, french, german, and spanish. *arXiv preprint arXiv:2312.04882*.

Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, and 1 others. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Yuxia Wang, Jonibek Mansurov, Petar Ivanov, Jinyan Su, Artem Shelmanov, Akim Tsvigun, Chenxi Whitehouse, Osama Mohammed Afzal, Tarek Mahmoud, Toru Sasaki, and 1 others. 2024. M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1369–1407.

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.

Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending against neural fake news. *Advances in neural information processing systems*, 32.