NextGen AI & Computing Journal (NAC), Volume 1, Issue 1 | April-June 2026| pp. 74-101

https://scientia-nexus.org/index.php/nac

DOI: https://doi.org/10.5281/zenodo.20434861

An Open-Access Peer-Reviewed Journal

Multilingual Generative AI Framework for Urdu and Regional Language Understanding Using Large Language Models

Shaista Jabeen
Email: shaistajabeen450@gmail.com
Department of Computer Science & Information Technology, University of South Punjab (USP), Multan, Punjab, Pakistan

Maria Jafar
Email: mariajafar17sep@gmail.com
Department of Computer Science & Information Technology, University of South Punjab (USP), Multan, Punjab, Pakistan

Usman Shafeeq
Email: usman.shafeeq@kfueit.edu.pk
Department of Data Science & Artificial Intelligence, Khawaja Fareed University of Engineering & Information Technology (KFUEIT), Rahim Yar Khan, Punjab, Pakistan

Palwasha Urooj Ch
Email: palwashaurooj297@gmail.com
Department of Computer Science & Information Technology, University of South Punjab (USP), Multan, Punjab, Pakistan

Ghulam Muhy Ud Deen Raee
Email: gmraees@gmail.com
Department of Computer Science & Information Technology, University of South Punjab (USP), Multan, Punjab, Pakistan

Hamid Ghous
Email: hamidghous@usp.edu.pk
Department of Computer Science & Information Technology, University of South Punjab (USP), Multan, Punjab, Pakistan

Mubasher Hussain Malik
Email: mubasher@usp.edu.pk
Department of Computer Science & Information Technology, University of South Punjab (USP), Multan, Punjab, Pakistan

Received: May 5, 2026 Accepted: May 29, 2026

Abstract

The rapid advancement of Large Language Models (LLMs) has significantly enhanced automated text generation, while simultaneously increasing the difficulty of distinguishing AI-generated content from human-written text. This challenge is particularly critical for low-resource languages such as Urdu, where reliable AI-content detection systems remain limited. To address this gap, this study proposes a multilingual AI-generated text detection framework specifically designed for Urdu language processing. A balanced benchmark dataset containing 1,800 human-authored and 1,800 AI-generated Urdu texts was developed using outputs from GPT-4o mini, Gemini, and Kimi AI. Comprehensive linguistic and statistical analyses were performed using features such as vocabulary richness, Type-Token Ratio (TTR), character diversity, sentence variability, and N-gram patterns, with significance validated through T-tests and Mann–Whitney U tests. Three multilingual transformer architectures, namely mDeBERTa-v3-base, DistilBERT-multilingual, and XLM-RoBERTa-base, were fine-tuned and evaluated on the proposed dataset. Experimental results demonstrated that mDeBERTa-v3-base achieved the best performance, obtaining an F1-score of 91.29% and an accuracy of 91.26% on the test dataset. The findings confirm the effectiveness of multilingual transformer models for AI-generated Urdu text detection and highlight their potential for supporting academic integrity, misinformation prevention, and trustworthy NLP applications in underrepresented language communities.

Keywords: Artificial Intelligence; Large Language Models; Urdu NLP; AI-Generated Text Detection; Multilingual Transformers; mDeBERTa; XLM-RoBERTa; DistilBERT; Low-Resource Languages; Natural Language Processing

I. Introduction

The emergence of Large Language Models (LLMs) has revolutionized the field of Natural Language Processing (NLP), enabling machines to generate highly coherent, contextually relevant, and human-like text across a wide range of applications. Advanced generative systems such as GPT-4, Gemini, Claude, and other state-of-the-art transformer-based architectures have demonstrated remarkable capabilities in text generation, summarization, translation, and conversational interaction. While these developments have accelerated innovation in education, content creation, journalism, and digital communication, they have simultaneously introduced critical challenges related to authenticity, transparency, and trustworthiness in textual content [1].

One of the most pressing concerns associated with modern generative AI systems is the increasing difficulty of distinguishing between human-authored and AI-generated text. As contemporary LLMs continue to improve in fluency and contextual reasoning, AI-generated outputs are becoming progressively more indistinguishable from genuine human writing. This challenge has significant implications for academic integrity, misinformation detection, automated journalism, online propaganda, and digital content verification [2]. The issue becomes even more severe in low-resource languages such as Urdu and other regional South Asian languages, where the availability of annotated datasets, linguistic resources, and AI detection systems remains extremely limited [3].

Unlike high-resource languages such as English and Chinese, Urdu suffers from a lack of large-scale benchmark corpora, standardized preprocessing pipelines, and domain-specific transformer models. Furthermore, Urdu presents several linguistic and computational complexities, including the Nasta'liq writing system, inconsistent word segmentation, rich morphology, extensive code-mixing with English, and orthographic variations [4]. These characteristics significantly complicate NLP tasks and reduce the effectiveness of models originally developed for high-resource languages. Consequently, Urdu-speaking communities are particularly vulnerable to the misuse of generative AI technologies, including AI-assisted plagiarism, fake news generation, automated propaganda, and manipulated educational content [5].

Despite the rapid advancement of generative AI systems, existing AI-generated text detection frameworks remain predominantly focused on English-language datasets and earlier generations of language models [6]. Most currently available detectors were trained using outputs from models such as GPT-2 or GPT-3 and therefore struggle to generalize effectively to more sophisticated architectures including GPT-4, Gemini, Claude, and LLaMA-based systems [7]. Moreover, very limited research has explored multilingual AI-generated text detection for Urdu and related regional languages, creating a substantial research gap in this emerging area [8].

To address these limitations, this study proposes a Multilingual Generative AI Framework for Urdu and Regional Language Understanding Using Large Language Models [9]. The proposed framework introduces a robust AI-generated text detection system specifically designed for Urdu-language content while supporting multilingual adaptability for regional languages [10]. A custom benchmark dataset named the UHAT Dataset was developed, consisting of balanced human-written and AI-generated Urdu texts collected from multiple domains, including literature, journalism, encyclopedic entries, and educational content. AI-generated counterparts were produced using advanced LLMs such as GPT-4o Mini, Gemini, and Kimi AI to ensure linguistic diversity and realism [11].

To overcome transformer input-length limitations, a sliding-window chunking strategy was implemented for long documents, resulting in a total of 7,667 processed text chunks. The chunking methodology preserved contextual continuity while improving model exposure to long-form linguistic patterns. A detailed overview of the chunking process and dataset characteristics is presented in Table I [12].

Table I: Data Chunking Summary

Metric	Value
Original Texts (Total)	3600
Texts Requiring Chunking	1657 (46%)
Total Chunks Created	7667
Average Chunks per Text	2.13
Chunked Dataset Shape	(7667, 6) where 6 represents the feature dimensions
Label Distribution in Chunks	Human: 4231 | AI: 3436
Chunk Length Statistics (Characters)	Minimum: 45 | Maximum: 450 | Average: 412.76 | Standard Deviation: 69.59

The dataset was subsequently divided into training, validation, and test subsets while maintaining balanced class distributions to ensure statistically reliable model evaluation. The distribution of dataset partitions is summarized in Table II.

Table II: Dataset Split Distribution

Dataset Split	Total Samples	Human Samples	AI Samples
Training Set	6133	3384 (55.2%)	2749 (44.8%)
Validation Set	767	423 (55.1%)	344 (44.9%)
Test Set	767	424 (55.3%)	343 (44.7%)

NextGen AI & Computing Journal (NAC), Volume 1, Issue 1 | April-June 2026| pp. 74-101

https://scientia-nexus.org/index.php/nac
DOI: https://doi.org/10.5281/zenodo.20434861

An Open-Access Peer-Reviewed Journal

The proposed framework leverages multilingual transformer architectures, including mDeBERTa-v3-base, DistilBERT-multilingual, and XLM-RoBERTa-base, fine-tuned specifically for AI-generated Urdu text detection [13]. These models were selected due to their strong contextual representation capabilities and multilingual semantic understanding. To further enhance model robustness, several Urdu-specific preprocessing techniques were incorporated, including Unicode normalization, diacritic removal, whitespace normalization, and preservation of Urdu punctuation. These preprocessing strategies significantly improved textual consistency while reducing noise associated with low-resource language processing [14]. In addition to deep learning-based classification, this research also investigates linguistic and statistical characteristics distinguishing AI-generated and human-authored Urdu texts. Metrics such as vocabulary richness, average text length, unique word distributions, and chunk-level statistics were analyzed to provide interpretable insights into generative writing behavior [15]. A summary of these linguistic characteristics is presented in Table III.

Table III: Summary Statistics of Human and AI-Generated Texts

Metric	Human	AI
Total Texts	1,800	1,800
Total Words	356,276	297,056
Total Characters	1,599,753	1,346,400
Unique Words	18,877	11,628
Average Text Length	888.75	748.00
Average Words per Text	197.93	165.03
Vocabulary Richness	109.02	96.36

Experimental evaluation demonstrated that multilingual transformer models achieved strong classification performance on the proposed dataset. Among the evaluated architectures, mDeBERTa-v3-base achieved the highest overall performance, outperforming DistilBERT-multilingual and XLM-RoBERTa-base across accuracy, F1-score, and precision metrics [16]. Detailed training and evaluation results are reported in Table IV and Table V, respectively.

Table IV: Model Training Performance

Model	F1 Score	Accuracy	Precision
microsoft/mdeberta-v3-base	92.98%	92.96%	93.55%
distilbert/distilbert-base-multilingual-cased	89.47%	89.44%	90.11%
FacebookAI/xlm-roberta-base	90.89%	90.87%	91.82%

Table V: Evaluation on Test Set

Model	Accuracy	F1 Score	Precision
mDeBERTa-v3-base	0.9126	0.9129	0.9232
DistilBERT-multilingual	0.8957	0.8960	0.9016
XLM-RoBERTa-base	0.8905	0.8907	0.9033

The contributions of this research can be summarized as follows:

1. Development of the UHAT Dataset, a balanced benchmark dataset for Urdu AI-generated text detection.
2. Introduction of a multilingual transformer-based detection framework optimized for Urdu and regional language processing.
3. Integration of language-aware preprocessing and chunking strategies to improve contextual understanding in low-resource settings.
4. Comprehensive comparative evaluation of state-of-the-art multilingual transformer architectures.
5. Linguistic and statistical analysis of AI-generated versus human-written Urdu content.

Overall, this work establishes a scalable and adaptable framework for multilingual AI-generated text detection in low-resource languages. By bridging the gap between advanced generative AI systems and trustworthy content verification mechanisms, the proposed study contributes toward strengthening academic integrity, misinformation prevention, and responsible AI deployment within underrepresented linguistic communities [17].

NextGen AI & Computing Journal (NAC), Volume 1, Issue 1 | April-June 2026| pp. 74-101

https://scientia-nexus.org/index.php/nac
DOI: https://doi.org/10.5281/zenodo.20434861

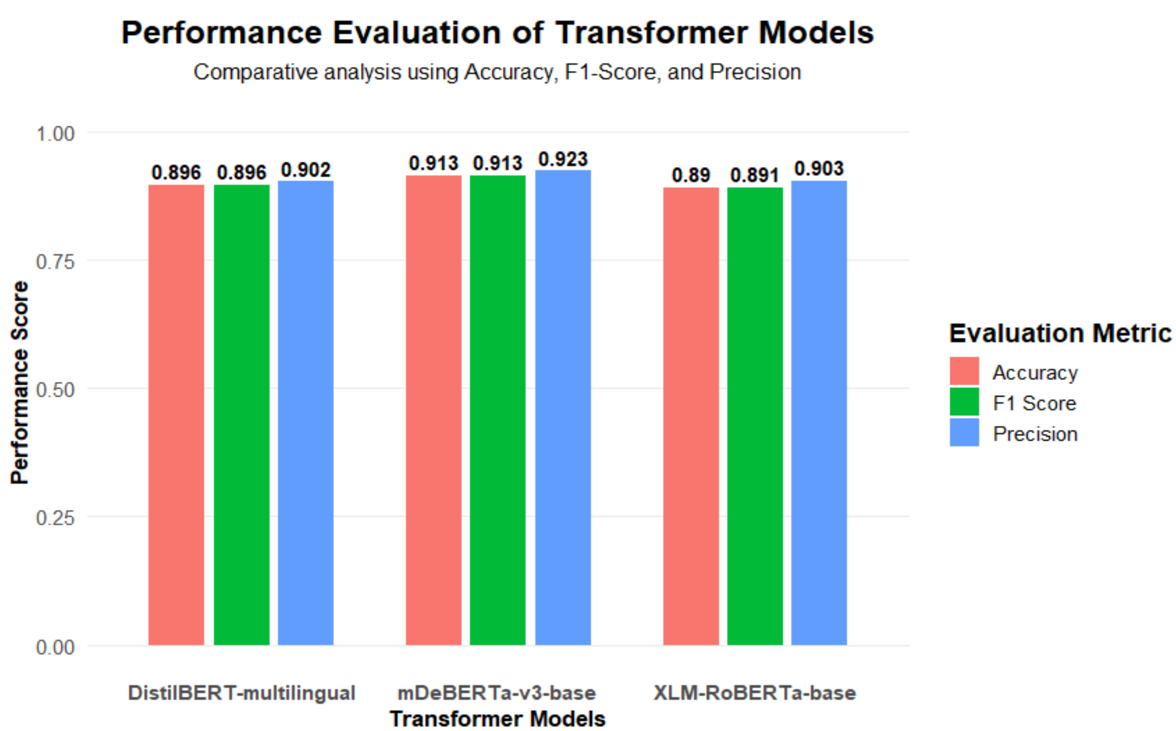
An Open-Access Peer-Reviewed Journal

IV. Methodology

The proposed framework introduces a comprehensive multilingual AI-based methodology for detecting AI-generated Urdu text using advanced transformer architectures, statistical linguistic analysis, and language-aware preprocessing strategies [17]. The methodology was specifically designed to address the challenges associated with low-resource languages such as Urdu, where limited annotated datasets, orthographic complexity, and linguistic variability significantly hinder the performance of conventional Natural Language Processing (NLP) systems [18]. The overall workflow of the proposed framework consists of four major stages:

1. Linguistic and Statistical Feature Analysis
2. Urdu-Specific Text Preprocessing
3. Dataset Chunking and Experimental Setup
4. Transformer-Based Model Fine-Tuning and Evaluation

A detailed overview of the methodological pipeline is illustrated in Figure 1.



NextGen AI & Computing Journal (NAC), Volume 1, Issue 1 | April-June 2026| pp. 74-101

https://scientia-nexus.org/index.php/nac
DOI: https://doi.org/10.5281/zenodo.20434861

An Open-Access Peer-Reviewed Journal

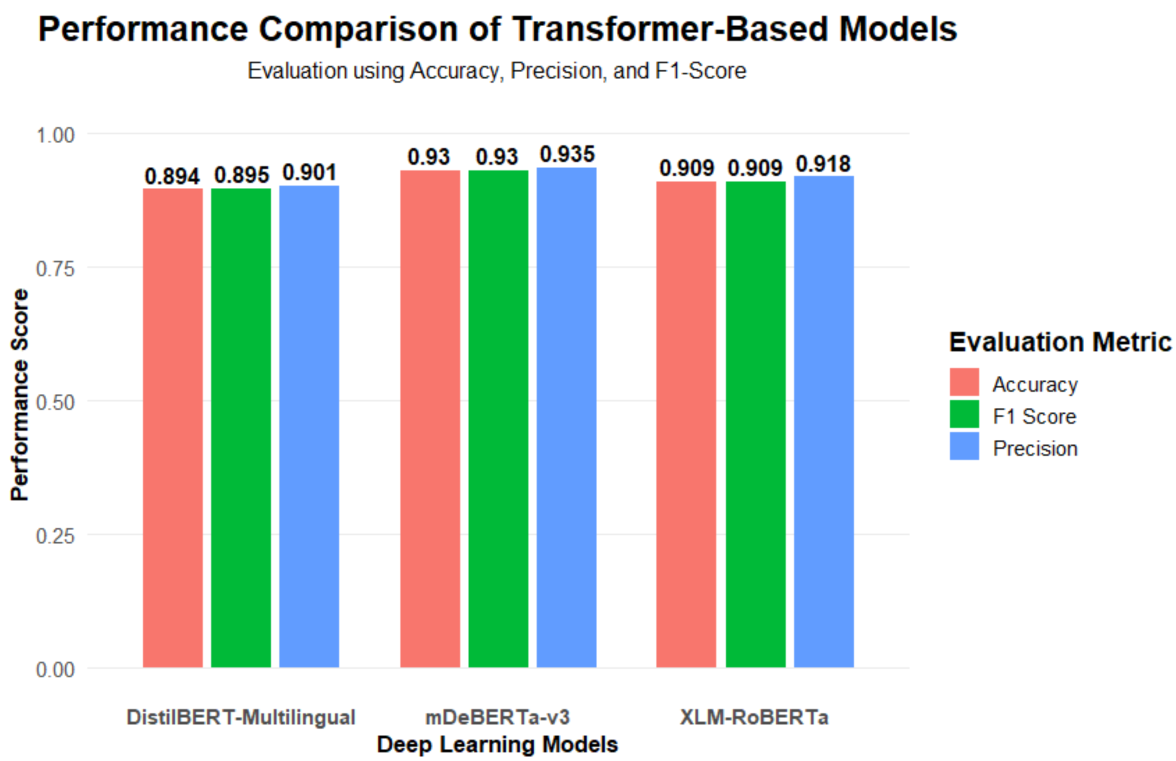


Figure 1: Proposed Multilingual AI Detection Framework

A. Linguistic and Statistical Analysis

To investigate the stylistic and structural differences between human-authored and AI-generated Urdu texts, an extensive linguistic and statistical analysis was conducted. This analysis aimed to identify interpretable textual patterns that could complement deep learning-based classification models and provide additional insight into the behavior of generative AI systems [19]. Several textual and linguistic features were extracted from each document, including:

a. Text length

b. Word count

c. Sentence count

d. Average word length

NextGen AI & Computing Journal (NAC), Volume 1, Issue 1 | April-June 2026| pp. 74-101
https://scientia-nexus.org/index.php/nac
DOI: https://doi.org/10.5281/zenodo.20434861
An Open-Access Peer-Reviewed Journal

e. Sentence length variability
f. Character diversity
g. Punctuation density
h. Vocabulary richness

Vocabulary richness was quantified using the Type-Token Ratio (TTR), which measures lexical diversity by calculating the ratio between unique words and total words within a text. In addition, N-gram analysis involving both bigrams and trigrams was performed to evaluate phrase repetition, contextual variation, and linguistic diversity across human and AI-generated samples [20]. To statistically validate the observed differences, both Independent Sample T-tests and Mann–Whitney U tests were applied across all extracted metrics. Experimental findings revealed statistically significant differences in five out of six linguistic complexity measures between human-written and AI-generated Urdu texts [21]. Human-authored texts demonstrated:

a. Higher vocabulary diversity
b. Greater sentence length variability
c. More dynamic linguistic structures
d. Increased lexical richness

Specifically, human-written texts achieved a higher average Type-Token Ratio (TTR) of 0.7091, compared with 0.6741 for AI-generated content. Similarly, the average word length in human text was marginally higher (3.5293) than that of AI-generated text (3.4809). Furthermore, sentence length variability was substantially greater in human-authored content (13.36) relative to AI-generated text (4.06), indicating that human writing tends to exhibit more natural structural variation [22]. In contrast, AI-generated texts produced slightly longer average sentences overall (152.17 words) compared to human-written texts (143.74 words), reflecting the tendency of generative models to produce more uniform and syntactically structured outputs. However, punctuation density showed no statistically significant difference between the two categories ($p = 0.4349$), suggesting that punctuation usage alone is insufficient for reliable AI-text discrimination. A summary of the linguistic complexity analysis is presented in Table VI.

https://scientia-nexus.org/index.php/nac
DOI: https://doi.org/10.5281/zenodo.20434861
An Open-Access Peer-Reviewed Journal

Table VI. Linguistic Complexity Analysis Between Human and AI-Generated Urdu Texts

Metric	**Human Text**	**AI-Generated Text**	**Observation**
Type-Token Ratio (TTR)	0.7091	0.6741	Human text exhibits greater lexical diversity
Average Word Length	3.5293	3.4809	Human text contains slightly longer words
Sentence Length Variability	13.36	4.06	Human writing shows higher structural variation
Average Sentence Length	143.74	152.17	AI text produces longer and more uniform sentences
Punctuation Density	Similar	Similar	No statistically significant difference
N-gram Diversity	Higher	Lower	AI text shows more repetitive phrase patterns

These findings provide important interpretable evidence supporting the deep learning models. Lower TTR values and reduced N-gram diversity in AI-generated content indicate repetitive or formulaic writing behavior, while reduced sentence variability reflects the structured nature of transformer-generated text. Consequently, these linguistic indicators help explain the semantic patterns learned by multilingual transformer architectures during classification [23].

NextGen AI & Computing Journal (NAC), Volume 1, Issue 1 | April-June 2026| pp. 74-101

https://scientia-nexus.org/index.php/nac
DOI: https://doi.org/10.5281/zenodo.20434861

An Open-Access Peer-Reviewed Journal

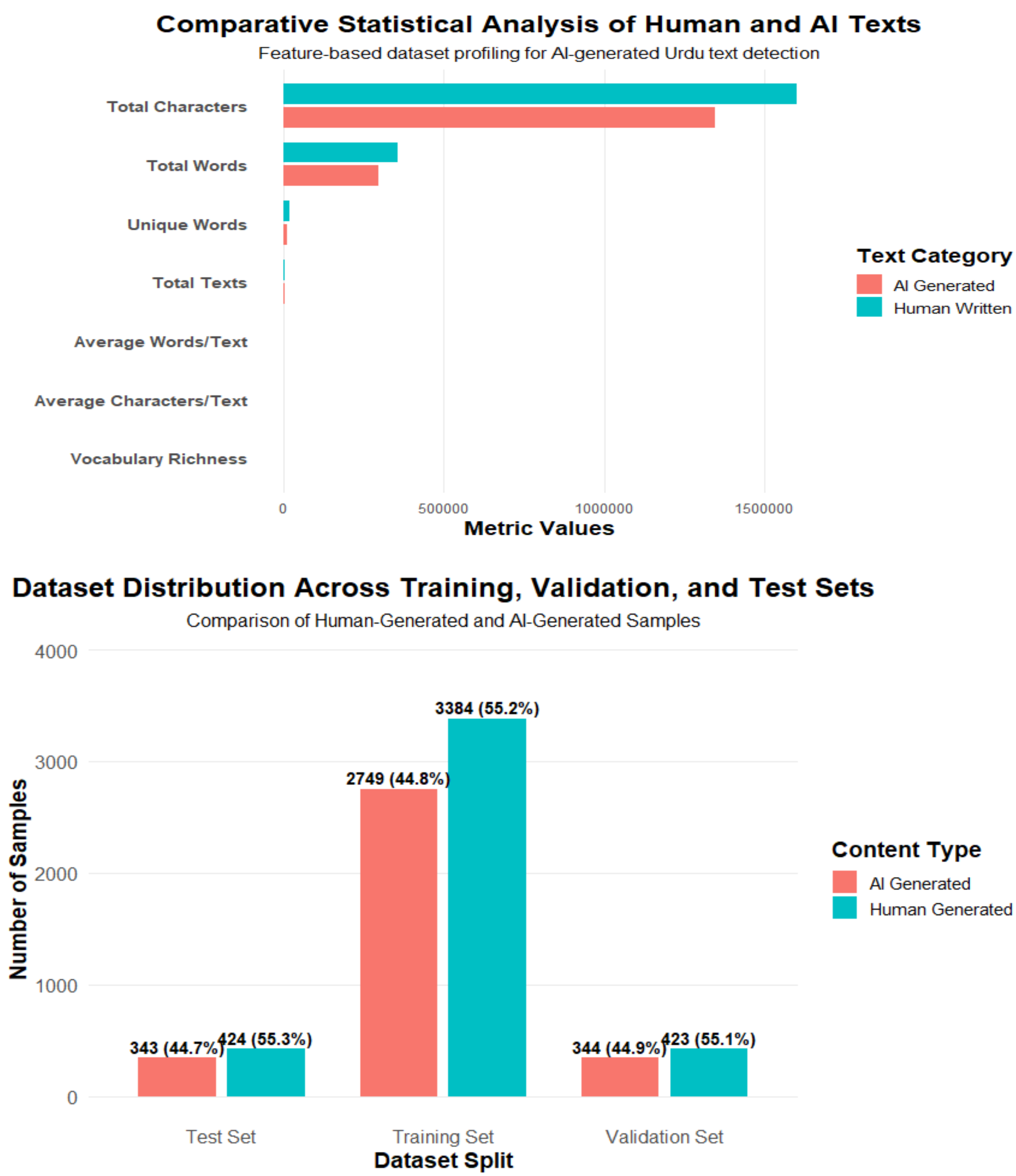


Figure 2: Comparative Linguistic Analysis of Human vs AI-Generated Urdu Texts

B. Urdu Text Preprocessing

Due to the linguistic and orthographic complexity of Urdu, a language-aware preprocessing pipeline was developed to optimize textual consistency and improve transformer model performance. Urdu presents several unique NLP challenges, including script variations, inconsistent spacing, optional diacritics, and code-mixed vocabulary. Therefore, standard preprocessing approaches designed for English are often insufficient [24]. The preprocessing framework incorporated the following steps:

1. Unicode Normalization

Unicode normalization was applied to ensure consistent representation of Urdu characters and eliminate encoding inconsistencies caused by multiple Unicode variants.

2. Whitespace and Newline Removal

Excessive whitespace, line breaks, and formatting artifacts were removed to reduce textual noise and improve sentence continuity.

3. Special Character Filtering

Non-linguistic symbols and irrelevant special characters were eliminated while preserving meaningful Urdu punctuation marks required for sentence boundary identification.

4. Diacritic Removal (Harakat)

Urdu diacritics (Harakat) were removed to reduce vocabulary sparsity and improve generalization across semantically equivalent words with different orthographic forms.

5. Token Standardization

Common token inconsistencies and spacing irregularities were normalized to improve compatibility with multilingual tokenizer architectures [25]. These preprocessing strategies significantly improved textual quality while reducing data sparsity in this low-resource language setting. Although diacritic removal may reduce certain phonetic distinctions, it provides a practical trade-off by simplifying vocabulary distributions and improving model robustness. A complete preprocessing workflow is illustrated in Figure 3.

NextGen AI & Computing Journal (NAC), Volume 1, Issue 1 | April-June 2026| pp. 74-101
https://scientia-nexus.org/index.php/nac
DOI: https://doi.org/10.5281/zenodo.20434861
An Open-Access Peer-Reviewed Journal

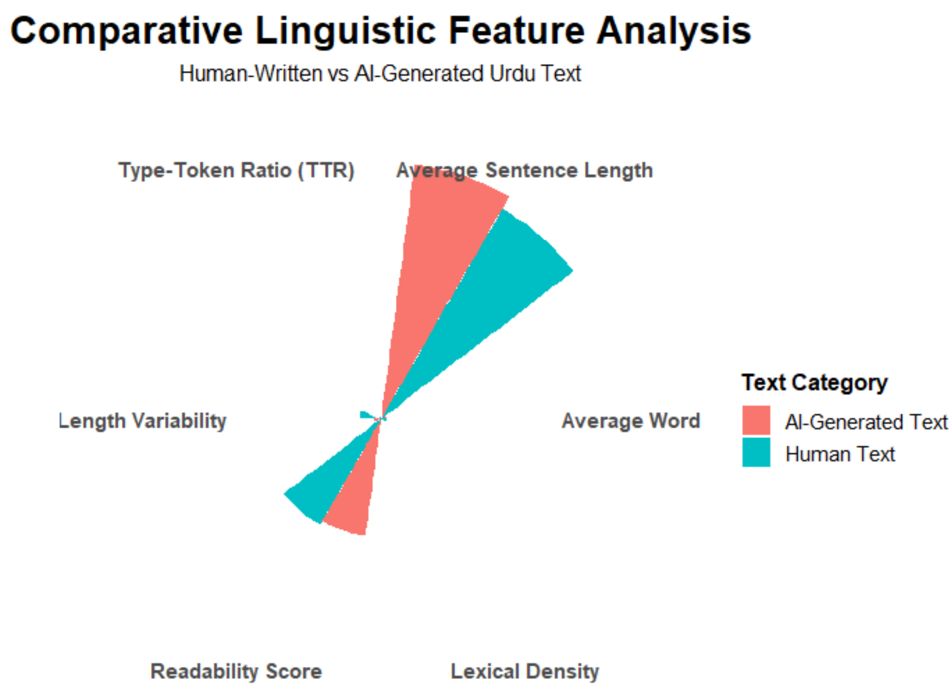


Figure 3: Urdu Text Preprocessing Pipeline

C. Experimental Setup

To ensure robust and statistically reliable model evaluation, the curated dataset was partitioned into:

- 80% Training Set
- 10% Validation Set
- 10% Test Set

Careful stratified sampling was performed to maintain balanced class distributions across all subsets and prevent class imbalance during model training. The dataset split distribution is summarized in Table VII.

Table VII. Dataset Split Configuration

Dataset Partition	Percentage	Purpose
Training Set	80%	Model learning and optimization
Validation Set	10%	Hyperparameter tuning and early stopping
Test Set	10%	Final performance evaluation

Sliding Window Chunking Strategy

Transformer architectures possess fixed input-length limitations, making it difficult to process long Urdu documents directly. To address this issue, a sliding-window chunking mechanism was implemented for all texts exceeding 450 characters. The chunking strategy:

a) Divided long texts into overlapping segments
b) Preserved semantic continuity between chunks
c) Increased dataset diversity
d) Improved contextual learning during training

This process expanded the dataset from 3,600 original documents to **7**,667 contextual chunks, substantially improving model exposure to diverse linguistic patterns. The chunking methodology is illustrated in Figure 4.

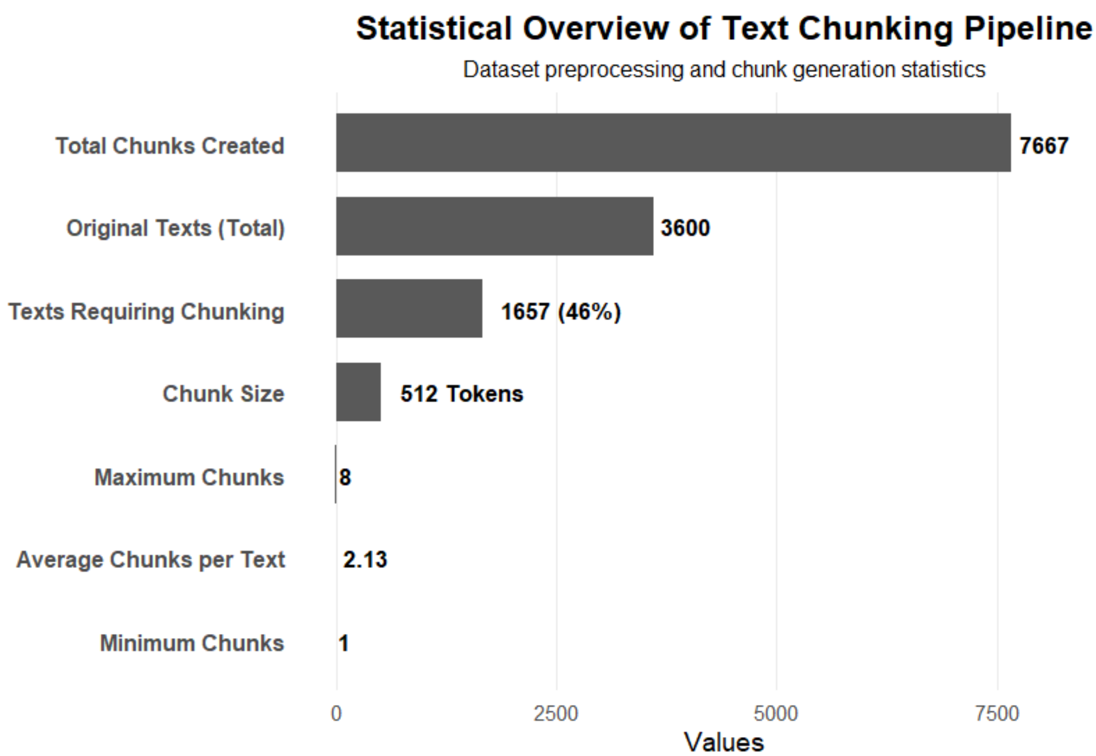


Figure 4: Sliding Window Chunking Strategy

D. Transformer-Based Model Fine-Tuning

Three state-of-the-art multilingual transformer architectures were selected for fine-tuning due to their strong multilingual semantic understanding and transfer-learning capabilities:

a) microsoft/mdeberta-v3-base

b) distilbert-base-multilingual-cased

c) FacebookAI/xlm-roberta-base

These architectures were chosen because they:

a) Support multilingual contextual embeddings

b) Perform effectively in low-resource language settings

c) Demonstrate strong cross-lingual transfer capabilities

d) Handle morphologically rich languages efficiently

The overall transformer training pipeline is illustrated in Figure 5.

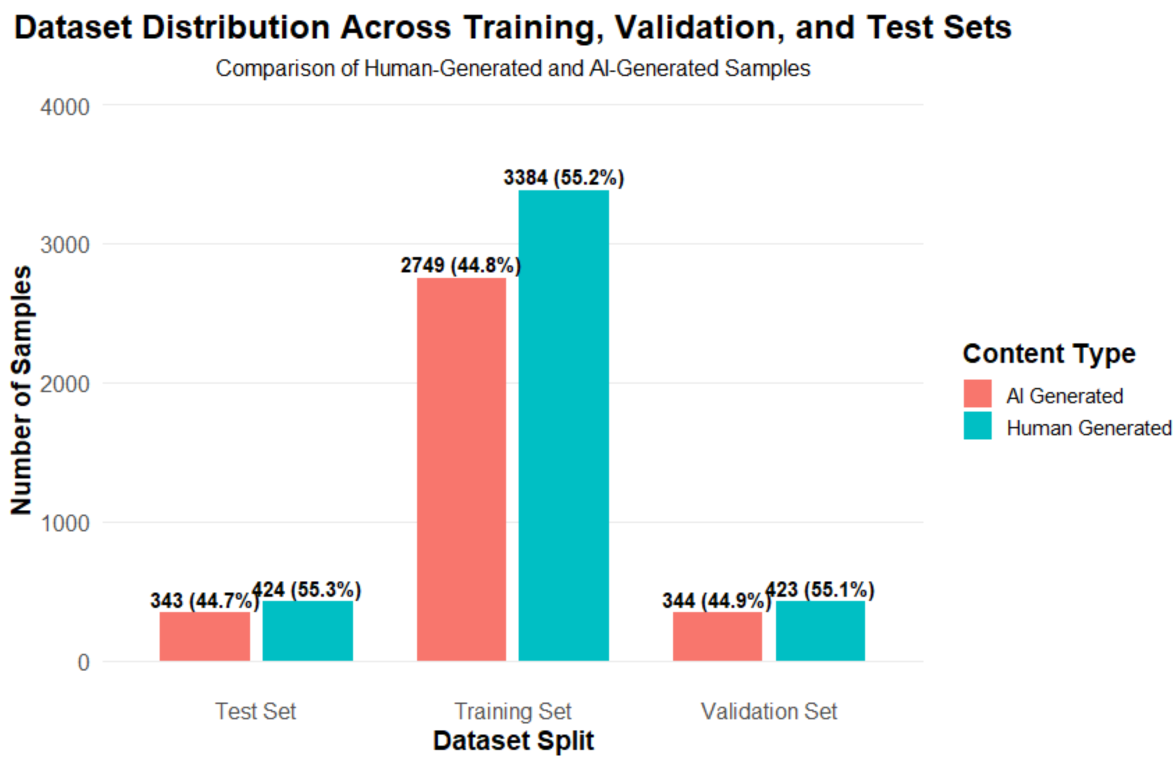


Figure 5: Transformer Model Fine-Tuning Workflow

E. Training Procedure and Optimization

To maximize model generalization and minimize overfitting, several optimization strategies were employed during training:

Early Stopping

Training was terminated automatically when validation loss ceased improving, thereby preventing overfitting.

Model Checkpointing

Best-performing model weights were preserved during training for optimal final evaluation.

Learning Rate Scheduling

Dynamic learning-rate adjustments were implemented to stabilize convergence during fine-tuning.

AdamW Optimizer

The AdamW optimization algorithm was used with carefully tuned learning rates to improve convergence stability and classification performance.

Regularization Techniques

Dropout and weight regularization were applied to improve model robustness on unseen Urdu text samples. The experimental configuration is summarized in Table VIII.

Table VIII. Model Training Configuration

Parameter	**Configuration**
Optimizer	AdamW
Training Strategy	Fine-tuning
Early Stopping	Enabled
Learning Rate Scheduling	Enabled
Chunking Threshold	450 Characters
Validation Monitoring	Validation Loss
Model Checkpointing	Enabled
Classification Task	Binary Classification

F. Evaluation Metrics

To comprehensively evaluate model performance, multiple classification metrics were employed:

a) Accuracy

NextGen AI & Computing Journal (NAC), Volume 1, Issue 1 | April-June 2026| pp. 74-101

https://scientia-nexus.org/index.php/nac
DOI: https://doi.org/10.5281/zenodo.20434861

An Open-Access Peer-Reviewed Journal

b) Precision

c) Recall

d) F1-Score

These metrics were selected to ensure balanced assessment of both detection sensitivity and classification reliability, particularly in multilingual low-resource environments.

The overall methodological framework combines statistical interpretability, Urdu-specific linguistic processing, and advanced transformer-based deep learning to establish a scalable and robust solution for multilingual AI-generated text detection [25]. The modular architecture further enables future adaptation to additional regional languages and evolving generative AI systems [26].

V. Results and Discussion

A. Model Training Summary

To evaluate the effectiveness of multilingual transformer architectures for AI-generated Urdu text detection, three state-of-the-art pre-trained models were fine-tuned using the curated UHAT dataset:

a) microsoft/mdeberta-v3-base

b) distilbert-base-multilingual-cased

c) FacebookAI/xlm-roberta-base

The models were trained using the training subset, while performance monitoring and hyperparameter optimization were conducted on the validation set. Early stopping and model checkpointing strategies were implemented to prevent overfitting and maximize generalization capability on unseen data. The AdamW optimizer, combined with adaptive learning-rate scheduling, ensured stable convergence throughout training [27]. The overall training performance of the evaluated models is summarized in Table IX.

Table IX. Training Performance of Multilingual Transformer Models

Model	F1 Score	Accuracy	Precision
microsoft/mdeberta-v3-base	92.98%	92.96%	93.55%
FacebookAI/xlm-roberta-base	90.89%	90.87%	91.82%
distilbert-base-multilingual-cased	89.47%	89.44%	90.11%

The results indicate that all three multilingual transformer architectures achieved strong classification performance despite the linguistic complexity and low-resource nature of Urdu. Among the evaluated models, mDeBERTa-v3-base consistently demonstrated the highest performance across all evaluation metrics, indicating its superior capability for contextual semantic understanding and multilingual representation learning [28].

B. Final Evaluation on Test Set

To assess the robustness and generalization capability of the proposed framework, the best-performing checkpoints from each model were evaluated on the independent test dataset. Performance was measured using Accuracy, Precision, Recall, and F1-score metrics to ensure balanced evaluation of classification quality [29]. The final test-set evaluation results are presented in Table X.

Table X. Final Evaluation Results on Test Dataset

Model	**Accuracy**	**Precision**	**F1 Score**
mDeBERTa-v3-base	91.26%	92.32%	91.29%
DistilBERT-Multilingual	89.57%	90.16%	89.60%
XLM-RoBERTa-base	89.05%	90.33%	89.07%

The experimental findings demonstrate that multilingual transformer models are highly effective for distinguishing between human-authored and AI-generated Urdu text. The mDeBERTa-v3-base model achieved the best overall performance, obtaining an accuracy of **91.26%**, precision of **92.32%**, and F1-score of **91.29%** on the test dataset. These results highlight the effectiveness of DeBERTa's disentangled attention mechanism and enhanced multilingual contextual encoding in capturing subtle linguistic variations present in Urdu text. Similarly, XLM-RoBERTa**-base** demonstrated strong multilingual generalization performance, achieving an F1-score of **89.07%** and precision of **90.33%**, confirming its robustness in multilingual NLP environments. Although DistilBERT-Multilingual produced slightly lower results compared to the other architectures, it still achieved competitive performance while offering the advantage of reduced computational complexity and faster inference speed. The consistently high precision and F1-scores across all evaluated models indicate balanced

NextGen AI & Computing Journal (NAC), Volume 1, Issue 1 | April-June 2026| pp. 74-101

https://scientia-nexus.org/index.php/nac
DOI: https://doi.org/10.5281/zenodo.20434861

An Open-Access Peer-Reviewed Journal

classification behavior for both AI-generated and human-written text categories. These findings further validate the effectiveness of transfer learning for low-resource language processing, demonstrating that multilingual transformer architectures can successfully generalize to Urdu despite its script complexity, morphological richness, and orthographic variability. The superior performance of mDeBERTa-v3-base suggests that newer multilingual transformer models possess stronger contextual representation capabilities, making them highly suitable for advanced NLP applications involving underrepresented regional languages.

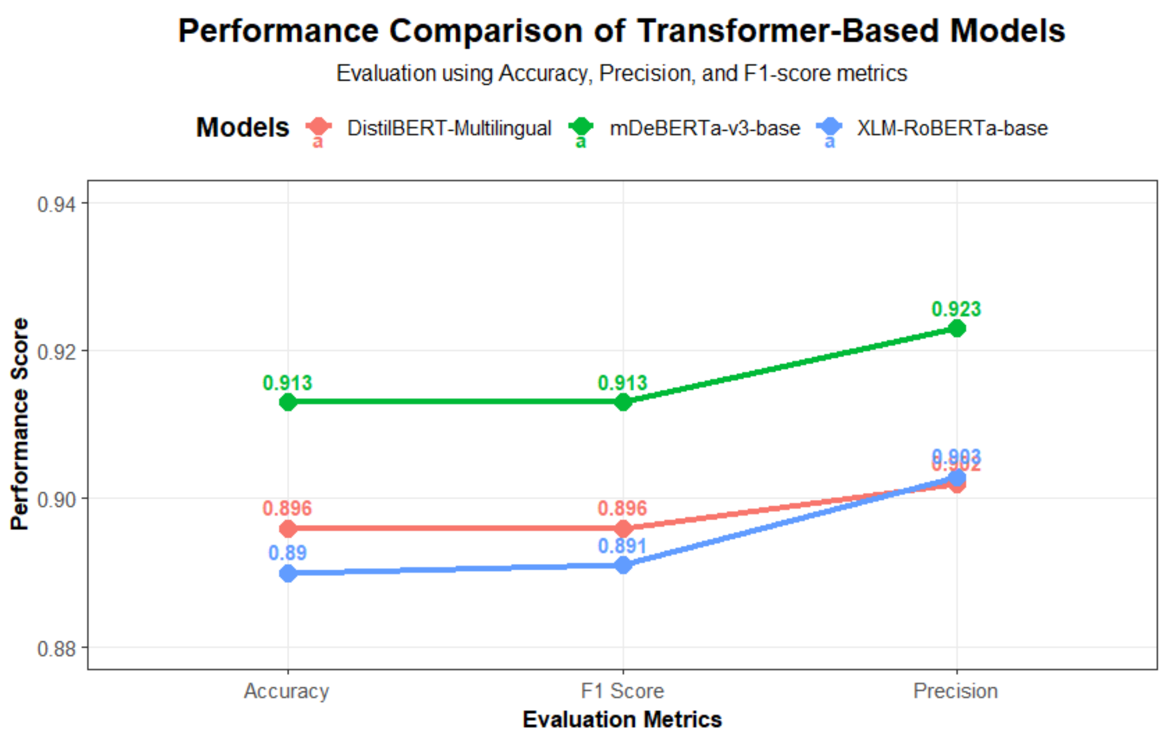


Figure 6: Comparative Performance Analysis of Transformer Models

C. Linguistic Interpretation of Results

Beyond numerical performance metrics, the obtained results also provide important linguistic insights into the structural behavior of AI-generated Urdu text. Statistical and linguistic analysis revealed that AI-generated content tends to exhibit:

a) Lower lexical diversity

b) Reduced sentence-length variability

c) More repetitive N-gram patterns

d) More structurally uniform writing styles

In contrast, human-authored Urdu text demonstrated:

a) Greater vocabulary richness

b) Higher contextual variation

c) More dynamic sentence structures

d) Increased stylistic diversity

These differences explain why multilingual transformer models were able to distinguish AI-generated text effectively. Models such as mDeBERTa-v3-base successfully captured semantic consistency, contextual coherence, and hidden linguistic patterns associated with machine-generated writing. The linguistic findings also complement the deep learning framework by providing interpretable evidence regarding the features learned during classification. A visual comparison of linguistic complexity measures is presented in **Figure 7**.

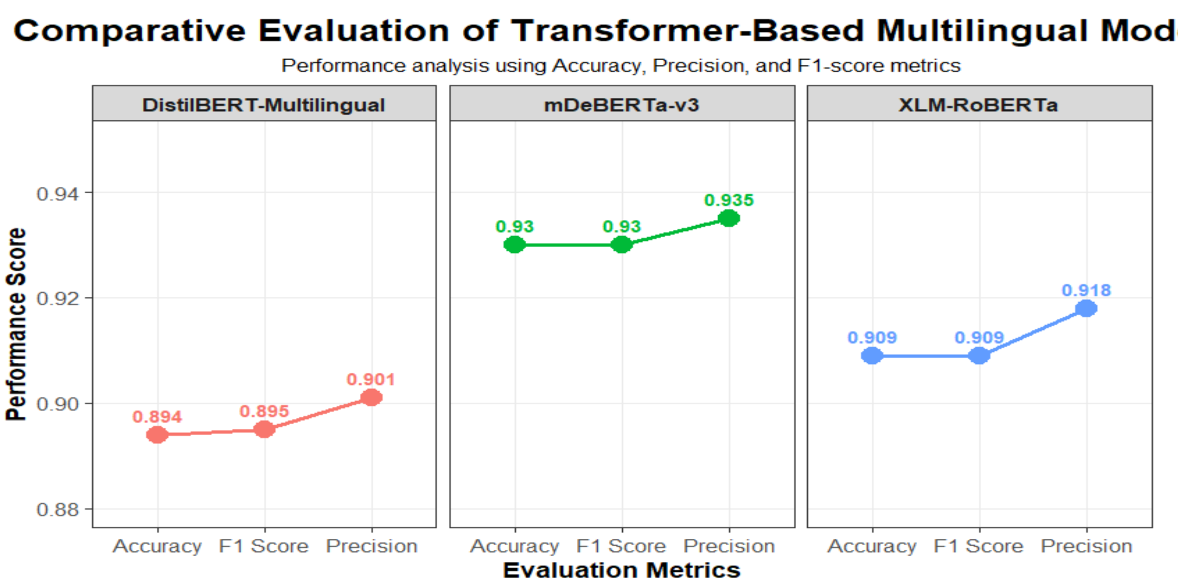


Figure 7: Linguistic Feature Comparison Between Human and AI-Generated Urdu Texts

D. Cross-Model Comparative Discussion

A comparative analysis of the evaluated architectures reveals several important observations:

1. mDeBERTa-v3-base

The model achieved the highest overall performance due to its advanced disentangled attention mechanism and enhanced positional encoding strategy. Its architecture enables more precise contextual understanding, which is particularly valuable for morphologically rich languages such as Urdu.

2. XLM-RoBERTa-base

XLM-RoBERTa-base exhibited strong multilingual robustness and semantic representation capabilities. The model demonstrated high stability across validation and testing phases, confirming its effectiveness for multilingual low-resource NLP tasks.

3. DistilBERT-Multilingual

Although slightly less accurate than the other architectures, DistilBERT-Multilingual provided competitive results while significantly reducing computational cost and inference time. This makes it a practical choice for lightweight or resource-constrained deployment environments.

The comparative performance analysis demonstrates that transformer-based multilingual transfer learning is highly effective for AI-generated text detection in regional languages.

E. Implications for Combating Generative AI Misuse

The successful development of an Urdu AI-generated text detection framework carries substantial real-world implications for both academia and digital media ecosystems.

Academic Integrity

In educational settings, generative AI systems are increasingly being used to produce essays, assignments, and academic reports. The proposed framework provides institutions and educators with an automated mechanism for verifying the authenticity of submitted Urdu-language content, thereby strengthening academic integrity and reducing AI-assisted plagiarism.

Misinformation Detection

The framework also offers significant potential for combating misinformation and AI-generated propaganda in Urdu-language digital media. Generative AI can rapidly produce misleading narratives, fake news, and manipulated political content, particularly in languages lacking effective moderation systems.

By accurately identifying AI-generated text, the proposed system can support:

a) Journalists
b) Fact-checkers
c) Social media moderators
d) Educational institutions
e) Government regulatory bodies

in detecting and limiting the spread of deceptive AI-generated information.

Low-Resource Language Advancement

Most existing AI detection systems are heavily focused on English-language datasets and high-resource languages. This research addresses a critical gap by providing one of the first large-scale multilingual AI detection frameworks specifically optimized for Urdu and regional South Asian languages. Consequently, this work contributes not only to the advancement of multilingual NLP research but also to the responsible deployment of generative AI technologies in underrepresented linguistic communities.

F. Overall Findings

The proposed multilingual AI framework successfully demonstrates that transformer-based architectures can effectively distinguish AI-generated Urdu text from human-authored content with high reliability and scalability. The integration of:

a) Urdu-specific preprocessing
b) Linguistic feature analysis
c) Sliding-window chunking
d) Multilingual transformer fine-tuning

collectively contributed to strong classification performance across all evaluated models. Overall, the findings establish a strong foundation for future research in multilingual AI-content detection, low-resource NLP, and trustworthy generative AI governance for regional languages.

VI. Conclusion

This study presents a robust and effective framework for detecting AI-generated Urdu text, addressing a critical gap in low-resource language processing amid the rapid rise of generative AI technologies. A balanced benchmark dataset comprising human-written and AI-generated Urdu texts was developed and analyzed using both linguistic-statistical techniques and advanced multilingual transformer architectures. Experimental evaluation demonstrated that multilingual transformer models can successfully distinguish AI-generated Urdu content with high reliability. Among the evaluated models, **microsoft/mdeberta-v3-base** achieved the best overall performance, obtaining an F1-score of **91.29%**, precision of **92.32%**, and accuracy of **91.26%**. DistilBERT-Multilingual and XLM-RoBERTa-base also produced strong and competitive results, confirming the effectiveness of transfer learning for low-resource multilingual NLP tasks. The findings further reveal significant linguistic differences between human-authored and AI-generated Urdu text, particularly in vocabulary richness, sentence variability, and contextual diversity. These insights strengthen the interpretability of the proposed framework while enhancing detection reliability. Overall, this research establishes a strong foundation for multilingual AI-generated text detection in Urdu and regional languages. Beyond its technical contributions, the proposed framework offers important real-world applications in academic integrity, misinformation detection, digital content verification, and responsible AI governance for underrepresented linguistic communities.

REFERENCES

[1] I. Solaiman et al., "Release Strategies and the Social Impacts of Language Models," arXiv preprint arXiv:1908.09203, 2019.

[2] T. Gehrmann, H. Strobelt, and A. M. Rush, "GLTR: Statistical Detection and Visualization of Generated Text," in ACL Demo, 2019.

[3] J. Carlini et al., "Evaluating and Testing Unsupervised Models of Text Generation," arXiv preprint arXiv:2205.11933, 2022.

[4] E. Kirchenbauer et al., "A Watermark for Large Language Models," arXiv preprint arXiv:2301.10226, 2023.

[5] A. Conneau et al., "Unsupervised Cross-lingual Representation Learning at Scale," in Proc. ACL, 2020.

[6] M. S. Imran et al., "A survey of challenges and applications of natural language processing for under-resourced languages," ACM Comput. Surv. (CSUR), vol. 54, no. 4, pp. 1–38, 2022.

[7] R. Zahur, M. I. Khan, and A. Farooq, "Challenges in NLP for Urdu and Its Applications in Social Media Analysis," Int. J. Comput. Sci. Netw. Secur., vol. 22, no. 3, pp. 1–8, 2022.

[8] M. Z. Asghar et al., "Fake News Detection in Urdu Language Using Machine and Deep Learning Models," IEEE Access, vol. 9, pp. 103051–103065, 2021.

[9] S. Bashir et al., "Developing a corpus for fake news detection in the Urdu language," J. King Saud Univ. – Comput. Inf. Sci., 2022.

[10] S. Ruder, I. Vulić, and A. Søgaard, "A Survey of Cross-Lingual Word Embedding Models," J. Artif. Intell. Res., vol. 65, pp. 569–631, 2019.

[11] P. He et al., "DeBERTa: Decoding-enhanced BERT with Disentangled Attention," arXiv preprint arXiv:2006.03654, 2020.

[12] Y. Liu et al., "Multilingual BERT: Effective Pretraining for Low-resource Languages," arXiv preprint arXiv:1901.07291, 2019.

[13] A. Conneau et al., "XLM-R: Robust Cross-lingual Representation Learning at Scale," arXiv preprint arXiv:1911.02116, 2019.

[14] S. Malmasi et al., "A Study of Stylometric and Lexical Features for Identifying Machine Generated Text," in Proc. NAACL-HLT, 2020.

[15] P. Potthast et al., "A Stylometric Inquiry into Hyperpartisan and Fake News," in Proc. ACL, 2018.

[16] M. Dou et al., "GPT detectors are biased against non-native English writers," arXiv preprint arXiv:2304.02819, 2023.

[17] R. Brown et al., "DetectGPT: Zero-shot Detection of Generated Text via Probability Curvature," arXiv preprint arXiv:2301.11305, 2023.

[18] J. Kirchner, "How ChatGPT Hijacks the Essay," The Atlantic, Jan. 2023.

[19] A. Ali and H. Mehmood, "Fake news detection in low-resource languages: A case study on Urdu," arXiv preprint arXiv:2302.08754, 2023.

[20] Muhammad Ammar (2025). Urdu Human and AI text Dataset (UHAT). IEEE Dataport. https://dx.doi.org/10.21227/y77y-9917

[1] M. C. Johnson, P. Patel, A. Ayers, and K. M. Spears, "Resource Management Challenges in Rural Dermatological Care: A Mapping Review," *Cureus*, vol. 17, no. 1, Jan. 2025, doi: 10.7759/cureus.77544.

[2] F. Basholli, M. R. Hayal, E. E. Elsayed, and D. A. Juraev, "Deep Learning for Skin Disease Classification: A Comparative Study of CNN and CNN-LSTM Architectures," *J. Comput. Data Technol.*, vol. 1, no. 1, pp. 40–49, 2025, doi: 10.71426/jcdt.v1.i1.pp40-49.

[3] G. Rehman, H. Shahab, A. Maqbool, and S. Hussain, "DEVELOPMENT OF AN IOT-BASED REAL-TIME PATIENT HEALTH MONITORING SYSTEM," *Pakistan J. Sci. Res.*, vol. 5, no. 02, pp. 170–174, 2025.

[4] B. Cassidy, C. Kendrick, A. Brodzicki, J. Jaworek-Korjakowska, and M. H. Yap, "Analysis of the ISIC image datasets: Usage, benchmarks and recommendations," *Med. Image Anal.*, vol. 75, p. 102305, 2022, doi: 10.1016/j.media.2021.102305.

[5] F. S. Malik, M. H. Yousaf, H. A. Sial, and S. Viriri, "Exploring dermoscopic structures for melanoma lesions' classification," *Front. Big Data*, vol. 7, p. 1366312, 2024, doi: 10.3389/fdata.2024.1366312.

[6] S. M. Thwin and H. S. Park, "Skin Lesion Classification Using a Deep Ensemble Model," *Appl. Sci.*, vol. 14, no. 13, p. 5599, 2024, doi: 10.3390/app14135599.

[7] Y. Doğan and C. Özdemir, "Enhancing Skin Cancer Diagnosis through the Integration of Deep Learning and Machine Learning Approaches," *Bilişim Teknol. Derg.*, vol. 17, no. 4, pp. 339–347, 2024, doi: 10.17671/gazibtd.1484037.

[8] P. Hermosilla, R. Soto, E. Vega, C. Suazo, and J. Ponce, "Skin Cancer Detection and Classification Using Neural Network Algorithms: A Systematic Review," *Diagnostics*, vol. 14, no. 4, p. 454, 2024, doi: 10.3390/diagnostics14040454.

[9] Z. R. Cai *et al.*, "Assessing the performance of artificial intelligence models in evaluating inflammatory skin disease severity: a systematic review and meta-analysis," *Br. J. Dermatol.*, vol. 193, no. 5, pp. 847–855, 2025, doi: 10.1093/bjd/ljaf250.

[10] A. Aboulmira *et al.*, "SkinHealthMate app: An AI-powered digital platform for skin disease diagnosis," *Syst. Soft Comput.*, vol. 6, p. 200166, 2024, doi: 10.1016/j.sasc.2024.200166.

[11] B. Ozdemir and I. Pacal, "A robust deep learning framework for multiclass skin cancer classification," *Sci. Rep.*, vol. 15, no. 1, p. 4938, 2025, doi: 10.1038/s41598-025-89230-7.

[12] J. Mohan, A. Sivasubramanian, S. V., and V. Ravi, "Enhancing skin disease classification leveraging transformer-based deep learning architectures and explainable AI," *Comput. Biol. Med.*, vol. 190, p. 110007, 2025, doi: 10.1016/j.compbiomed.2025.110007.

[13] M. Arshad, M. A. Khan, N. A. Almujally, A. Alasiry, M. Marzougui, and Y. Nam, "Multiclass skin lesion classification and localziation from dermoscopic images using a novel network-level fused deep architecture and explainable artificial intelligence," *BMC Med. Inform. Decis. Mak.*, vol. 25, no. 1, p. 215, 2025, doi: 10.1186/s12911-025-03051-2.

[14] K. Nawaz *et al.*, "Skin cancer detection using dermoscopic images with convolutional neural network," *Sci. Rep.*, vol. 15, no. 1, p. 7252, Mar. 2025, doi: 10.1038/s41598-025-91446-6.

[15] S. Fatima, M. U. Akram, S. Mohammad, and S. Bin Ahmed, "Deep learning in dermatopathology: applications for skin disease diagnosis and classification," *Discov. Appl. Sci.*, vol. 7, no. 9, p. 1006, 2025, doi: 10.1007/s42452-025-07138-3.

[16] Abbas, M. A. (2025). Advanced Synthesis and Multifunctional Characterization of Neodymium-Doped $Ba_2NiCoFe_{28-x}O_{46}$ X-Type Hexagonal Ferrites: A Comprehensive Study of Structural, Morphological, and Electromagnetic Properties. *Sch Acad J Biosci*, *8*, 1213-1227.

[17] Abbas, M. A., Junaid, M. J. M., Rasool, M. S., & Mahar, J. (2025). Structural and NLO Properties of Novel Organic 4-Bromo-4-Nitrostilbene Crystal: Experimental and DFT Study. *International Research Journal of Management and Social Sciences*, *6*(4), 1-20.

[18] Abbas, M. A., Junaid, M. J. M., Rasool, M. S., & Mahar, J. (2025). Structural and NLO Properties of Novel Organic 4-Bromo-4-Nitrostilbene Crystal: Experimental and DFT Study. *International Research Journal of Management and Social Sciences*, *6*(4), 1-20.

[19] Atif, H. M., Shahzad, A., Khan, M. Z., Abbas, M. A., & Mahar, J. (2025). Design of Novel drug as Potential Anti-Prostate Cancer Activity: Thiophene Derivatives against prostate cancer cell line as therapeutic agents using Pharmacokinetics molecular docking and DFT studies. *Indus Journal of Bioscience Research*, *3*(6), 548-559.

[20] Abbas, M. A., Khan, M. Z., Atif, H. M., Shahzad, A., & Mahar, J. (2025). Computer-Aided Analysis of Oxino-bis-Pyrazolederivative as a Potential Breast Cancer Drug Based on DFT, Molecular Docking, and Pharmacokinetic Studies: Compared with the Standard Drug Tamoxifen. *Indus Journal of Bioscience Research*, *3*(6), 535-537.

[21] Abbas, M. A., & Rasool, M. S. (2026). Eco-Friendly Synthesis of Ag–Co3O4 Nanoparticles for Visible-Light Photocatalysis and DFT-Based Nonlinear Optical Investigation. *Chemical Technology and Engineering Applications*, *1*(1), 23-34.

[22] Rasool, M. S., Abbas, M. A., Khan, M. J., Mahar, J., & Khan, M. Z. IDENTIFICATION OF NATURAL EGFR TYROSINE KINASE INHIBITORS FROM CHENOPODIUM QUINOA WILLD. VIA COMBINATORIAL IN SILICO AND PHARMACOLOGICAL SCREENING.

[23] Akram, S., Abbas, M. A., Mahar, J., Rasool, M. S., & Junaid, M. INTERFACIAL DEFECT PASSIVATION AND PHOTOPHYSICAL ENGINEERING OF $CSPBCL_3$ QUANTUM DOTS VIA BISBENZIMIDAZOLIUM LIGANDS FOR ADVANCED ELECTRONIC DEVICES.

[24] Junaid, M., Rasool, M. S., Abbas, M. A., & Mahar, J. (2024). Formulation Development and Evaluation of a Bilayered Tablet Containing Dapagliflozin and Metformin. *Global Research Journal of Natural Science and Technology*, *2*(3).

[25] Amin, M., Abbas, M. A., Mahar, J., Shahzad, M. S., & Rasool, M. S. (2026). Phyto-Mediated Green Synthesis and Physicochemical Characterization of Titanium Dioxide Nanoparticles for Environmental and Pharmacological Applications. Journal of Physical and Chemical Studies (JPCS), 1(4), 17–56. https://doi.org/10.5281/zenodo.19767807

[26] Abbas, M. A., Mahar, J., Ali, N., Junaid, M., & Rasool, M. S. (2026). Green Synthesis of SnO_2 Nanomaterials: Photocatalytic Degradation of Methylene Blue and DFT-Based Investigation of Nonlinear Optical Properties. Journal of Physical and Chemical Studies (JPCS), 1(3), 1–29. https://doi.org/10.5281/zenodo.19693725

[27] Abbas, M. A., Mahar, J., Ali, N., Junaid, M., & Rasool, M. S. (2026). Photocatalytic Dynamics of Organic Dye Degradation on Graphitic Carbon Nitride: An Integrated Experimental and Theoretical Investigation. Journal of Physical and Chemical Studies (JPCS), 1(2), 1–23. https://doi.org/10.5281/zenodo.19693515

[28] Abbas, M. A., Mahar, J., Ali, N., Junaid, M., & Rasool, M. S. (2026). Interfacial Defect Passivation and Photophysical Modulation in Cesium Lead Chloride Perovskite Quantum Dots Using Bisbenzimidazolium Ligands for Advanced Optoelectronic Devices. Journal of Physical and Chemical Studies (JPCS), 1(1), 1–18. https://doi.org/10.5281/zenodo.19666800

[29] Akram, S., Abbas, M. A., Mahar, J., Rasool, M. S., & Junaid, M. (2026). SYNTHESIS AND CHARACTERIZATION OF ZINC-DOPED CARBON DOTS FOR ENHANCED FLUORESCENCE APPLICATIONS. *Policy Research Journal, 4*(2), 168–177. https://policyrj.com/1/article/view/1550