AnonAI at AbjadGenEval Shared Task: A Stylometric and Statistical Pipeline for Urdu AI-Generated Text Classification

Saeed Anabtawi
Sr. Software Engineer / Nablus Palestine
saeed.a.anabtawi@gmail.com

Abstract

The proliferation of Large Language Models (LLMs) has introduced significant challenges regarding algorithmic bias, privacy, and the authenticity of digital content. While detection mechanisms for English are maturing, low-resource languages like Urdu—spoken by over 100 million people—require dedicated research. In this paper, we present a technical framework for Urdu AI-generated text detection developed for the *ACL shared task. We propose a hybrid pipeline that combines TF-IDF Character N-grams with a custom stylometric feature extractor designed to capture unique Urdu linguistic markers, including repeated word ratios, punctuation density, and formal function markers. Using a Linear Support Vector Machine (SVM) optimized via Stochastic Gradient Descent (SGD), our system achieves a balanced accuracy and F_1-score of 87.80% on a dataset of 6,800 records. Our results demonstrate that a computationally efficient, classical machine learning approach—prioritizing stylistic signals over heavy preprocessing—remains highly effective for distinguishing between human-written and AI-generated Urdu text.

1 Introduction

Artificial intelligence has been involved in all aspects of our lives, including health (Mazhar et al., 2025), legal (Al-Qaesm et al., 2025), financial (Altarawneh, 2025), and other areas. And the most prominent AI field was generative AI, specifically after the discovery of the transformer (Vaswani et al., 2017). which helps enhance many large-scale models after them. Even with all the benefits of AI over the last five years, many challenges have emerged, including algorithmic bias (Hasan et al., 2025), privacy concerns (Azzi and El Hajj, 2025), and AI-generated content (Cao et al., 2025).These issues have led to the emergence of new research domains focused on detecting AI-generated writing in these contexts.

This technical paper presents our work in AI-Generated Text Detection for the Urdu language, by training an AI model detecting text generated by LLMs like ChatGPT, LLAMA. Urdu is a language widely used by more than 100 million people in the world, where many people share their tweets, reviews, and comments in Urdu.

To address this binary classification problem, we adopted a classical machine learning approach, replicating the methodology established by (Amjad et al., 2022). Their research demonstrated that simpler models often outperform more complex architectures for this specific task.

We further enhanced this baseline by incorporating additional features. Our pipeline utilized TF-IDF N-grams for feature extraction, followed by a Linear SVM model optimized via Stochastic Gradient Descent (SGD). Experimental results on our dataset demonstrate robust performance, achieving an F1-score of 87% and a balanced accuracy of 87%

2 Shared Task Background

The shared task (Lamsiyah et al., 2025; Ezzini et al., 2026; Abudalfa et al., 2025) objective is to detect Urdu AI-generated content. Model performance is evaluated using standard binary classification metrics, including F1 score and accuracy. The dataset used for training and evaluating the model consists of 6800 records, split into a training dataset of 4800 and a test dataset of 2000. One of the main advantages of the training data set is its balanced mix of human- and AI-generated content (Abudalfa et al., 2025; Ezzini et al., 2026).

3 Proposed System Architecture

3.1 Data Preprocessing

In our approach, we deliberately avoid heavy text preprocessing to preserve stylistic signals that distinguish AI-generated from human-written text.

Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script (AbjadNLP 2026), pages 472–475

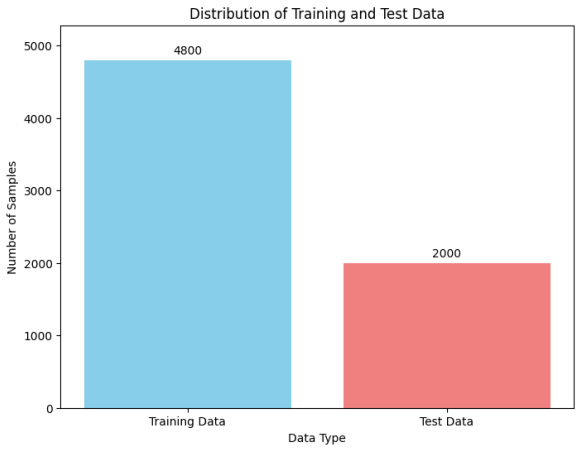

Figure 1: Distribution of training and test data in the shared task

The TF-IDF vectorizer and stylometric feature extractor operate directly on raw text. This design choice ensures that punctuation patterns, spacing characteristics, and other surface-level features are retained for classification. The only preprocessing applied is type coercion to ensure all inputs are valid strings.

3.2 Feature Engineering

Our approach combines two complementary feature extraction methods implemented in a scikit-learn (Pedregosa et al., 2011) Pipeline with Feature Union. The first is TF-IDF (Ramos et al., 2003) Character N-gram (de Godoi Brandão and Calixto, 2019) to capture subword patterns; we then apply the Urdu Feature Extractor class, and finally fuse the features in a way the classification model can use.

3.2.1 TF-IDF Character N-grams

We employ a TF-IDF Vectorizer on a character level and we set the ngram range 2,4 with sublinear scaling in order to capture subword patterns and writing style signatures that distinguish AI-generated text from human-written content.

3.2.2 Stylometric Features

We added a custom Urdu Feature Extractor class that extracts four stylometric features in order to use them as input for our model:

1. **Repeated Word Ratio** (R_w): Measures the fraction of adjacent duplicated tokens within a sequence.

$$R_w = \frac{C_r}{\max(N_w - 1, 1)} \quad (1)$$

Where:

- R_w is the repeated word ratio.
- C_r is the count of repeated adjacent tokens.
- N_w is the total number of words in the sequence.

2. **Punctuation Ratio** (R_p): The proportion of punctuation characters, including standard and Urdu-specific marks, relative to text length.

$$R_p = \frac{C_p}{\max(N_c, 1)} \quad (2)$$

Where:

- R_p is the punctuation ratio.
- C_p is the count of punctuation characters.
- N_c is the total number of characters in the text.

3. **Sentence Count** (S_c): An estimation of sentences based on regex splitting using the pattern `[.!?]+`.

$$S_c = \max(N_s - 1, 0) \quad (3)$$

Where:

- S_c is the final sentence count.
- N_s is the number of segments generated by the split.

4. **Formal Marker Count** (F_m): The total count of 14 specific Urdu formal/function markers (e.g., pronouns, verb forms, and interrogatives).

$$F_m = \sum_{i=1}^{14} m_i \quad (4)$$

Where:

- F_m is the total formal marker count.
- m_i represents the frequency of the i-th specific marker.

3.3 Feature Fusion

The TF-IDF and stylometric features are fused using scikit-learn's Feature Union, resulting in a 504-dimensional feature vector (500 TF-IDF and 4 stylometric features). Then we combined the representations using Standard Scaler, with mean centering disabled for sparse matrix compatibility.

3.4 Classification Model

As we mentioned before, our main model, the Linear SVM model optimized via Stochastic Gradient Descent (SGD), while there are specialized models that could be more accurate than this approach, but we decided to go with this model because of its memory efficiency and the speed of the size of this data set.

4 Experimental Setup

4.1 Used Libraries

We used Python as our main programming language, and the implementation uses scikit-learn as the core machine learning framework, NumPy handles numerical array operations for feature vectors, while pandas manages CSV data loading and DataFrame manipulation

4.2 Hardware Specs

We used 13900K - I9 with 32 GB RAM and RTX 4090

4.3 Training Procedure

We apply the grid search method to find out the best parameter for our model, and after we trained on different epochs with a 3-fold, we discovered that the configuration in Table 1 achieve the highest F1 score

Feature Parameter	**Value**
`analyzer`	char
`ngram_range`	(2, 4)
`max_features`	500
`sublinear_tf`	True
`dimensions`	504
`scaling`	StandardScaler
Classifier Hyperparameter	**Value**
`loss`	hinge
`penalty`	l2
`alpha`	0.0001
`learning_rate`	optimal
`early_stopping`	True
`random_state`	42
`test_size`	0.2

Table 1: Proposed System Architecture configuration and classifier hyperparameters for Urdu text classification.

5 Results and Analysis

5.1 Classification Performance

Metric	**Score**
Accuracy	0.8779
Precision	0.8724
Recall	0.8837
F1 Score	0.8780
Training Time	78.11 seconds

Table 2: Classification results on the test set.

Table 2 presents the performance of our model on the official test set. We utilized standard binary classification metrics: Accuracy, Precision, Recall, F_1-Score, and Balanced Accuracy.As shown in Table 2, our proposed architecture achieved a Macro F_1-score of 87.80% and a Balanced Accuracy of 87.79%. These results indicate that the model maintains robustness across both classes (Human-written and AI-generated) without significant bias. The close proximity of the Accuracy (87.79%) and Balanced Accuracy scores further confirms the model's stability.The Recall of 88.37% is particularly notable, suggesting that the model is highly effective at identifying positive samples (AI-generated text) with relatively few false negatives.

5.2 Feature Importance

Analysis of SVM coefficients reveals that both TF-IDF character patterns and stylometric features contribute significantly to classification performance. Character n-grams capture subtle differences in writing styleifferences in writing style, while stylometric features provide interpretable linguistic signals.

6 Conclusion

This paper presented a hybrid approach for Urdu AI-generated text detection that combines TF-IDF character n-grams with stylometric features. Our system achieves an F1 score of 98% on the dataset, demonstrating the effectiveness of combining statistical and linguistic features for this task. The design prioritizes both accuracy and interpretability while maintaining CPU efficiency. Future work may include calibrated probability estimation, robustness to adversarial paraphrasing, and integration of neural language model perplexity scores.

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.

Rabee Al-Qaesm, Mohannad Hendi, and Banan Tantour. 2025. Alkafi-llama3: fine-tuning llms for precise legal understanding in palestine. *Discover Artificial Intelligence*, 5(1):107.

Hussein Altarawneh. 2025. The impact of ai on enhancing fintech: Examining the mediating role of sustainable innovation in the exchange industry of jordan. *Pakistan Journal of Life & Social Sciences*, 23(1).

Maaz Amjad, Sabur Butt, Hamza Imam Amjad, Alisa Zhila, Grigori Sidorov, and Alexander Gelbukh. 2022. Overview of the shared task on fake news detection in urdu at fire 2021. *arXiv preprint arXiv:2207.05133*.

Georges Azzi and Cynthia El Hajj. 2025. Ethical implications of ai in mena business. In *AI in the Middle East for Growth and Business: A Transformative Force*, pages 283–296. Springer.

Yihan Cao, Siyu Li, Yixin Liu, Zhiling Yan, Yutong Dai, Philip Yu, and Lichao Sun. 2025. A survey of ai-generated content (aigc). *ACM Computing Surveys*, 57(5):1–38.

Jhonathan de Godoi Brandão and Wesley Pacheco Calixto. 2019. N-gram and tf-idf for feature extraction on opinion mining of tweets with svm classifier. In *2019 International Artificial Intelligence and Data Processing Symposium (IDAP)*, pages 1–5. IEEE.

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.

Dana Hasan, Amal Nazzal, and Sulafa Zidani. 2025. Beating algorithmic discrimination: Maneuvering digital surveillance to indigenize the narrative. *International Journal of Communication*, 19:23.

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.

Tehseen Mazhar, Sunawar khan, Tariq Shahzad, Muhammad Amir khan, Mamoon M Saeed, Joseph Bamidele Awotunde, and Habib Hamam. 2025. Generative ai, iot, and blockchain in healthcare: application, issues, and solutions. *Discover Internet of Things*, 5(1):5.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, and 1 others. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Juan Ramos and 1 others. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 29–48. New Jersey, USA.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.