iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** II **Month of publication:** February 2026

DOI: https://doi.org/10.22214/ijraset.2026.77702

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue II Feb 2026- Available at www.ijraset.com

Robust Detection of Paraphrased AI-Generated Text Using Deep Recurrent Neural Networks

Mohit Shirpurkar
School of Engineering and Technology, Pimpri Chinchwad University, 412106, Pune, Maharashtra

Abstract: ***The rapid advancement of large language models (LLMs) has made AI-generated text increasingly fluent and indistinguishable from human writing.***
However, malicious use of AI text for misinformation or plagiarism raises the need for reliable detectors. Simple detectors often fail when the AI-generated text is paraphrased by an adversary. In this work, we propose a robust detection framework based on deep recurrent neural networks (RNNs) that is resilient to paraphrasing. We compile a large dataset of AI-generated and human-written text (e.g., a 500K Kaggle corpus and simulate paraphrase attacks using state-of-the-art paraphrasing models. Our model employs a multi-layer Long Short-Term Memory (LSTM) network to capture sequential patterns and is trained with both original and paraphrased samples. In experiments, the proposed RNN classifier achieves high accuracy on unaltered AI text and retains strong performance on paraphrased adversarial examples (far exceeding the drop seen in baseline detectors. These results demonstrate that deep recurrent models, when properly trained, can detect AI-generated content even under paraphrasing attacks. We discuss implications for academic integrity and outline future enhancements such as multi-lingual extensions.
Keywords: Paraphrased AI-Generated Text, Deep Recurrent Neural Networks.

I. INTRODUCTION

The proliferation of AI-generated text (e.g. from models like GPT-4) has enabled numerous beneficial applications but also poses risks of misuse (e.g. misinformation, fake reviews, academic dishonesty). Detecting AI- generated content is therefore an active area of research.

However, straightforward detectors often assume a fixed distribution of AI text. Recent studies show that attackers can significantly degrade detector performance by paraphrasing the AI text. For example, Krishna et al. found that a paraphraser (DIPPER) reduced DetectGPT accuracy from 70.3.

Likewise, Sadasivan et al. demonstrated that recursive paraphrasing attacks cause sharp drops in detection rates for most existing method. These works highlight that detectors must be robust to paraphrasing strategies.

In this paper, we address the challenge of paraphrased AI-generated text. We develop a deep recurrent neural network (RNN) classifier (based on LSTM cells) trained on a mixture of original and paraphrased AI-generated texts along with human-written texts. By explicitly including paraphrase-augmented data in training, our model learns features that distinguish underlying AI-generated structures even after surface edits.

We validate our approach on large-scale datasets (e.g. the Kaggle 500K essay corpus and synthetically generated paraphrases) and compare with baselines. Our results show that the proposed RNN model maintains high detection accuracy even on paraphrased texts, significantly outperforming naive detectors. The contributions of this work are:

A demonstration that adversarial paraphrasing severely undermines standard AI-text detectors, motivating robust methods.

1) A novel LSTM-based detection framework trained on both original and paraphrased AI texts, capturing sequential linguistic patterns.
2) Experimental evaluation on publicly available datasets (e.g. Kaggle AI vs. Human) and adversarial paraphrases, showing that our model significantly reduces the attack-induced performance drop (cf. Table 1).
3) Analysis of feature patterns and ablation, and a discussion of future enhancements (multi-lingual data, hybrid models) to further improve robustness.

II. LITERATURE REVIEW

Detection of AI-generated text has been studied via statistical, linguistic, and machine learning methods. Early approaches used stylometric features (e.g. part-of-speech frequencies, function word usage) with classifiers. For example, Huang et al. note that stylometric approaches achieved high accuracy on student essays using features like TF-IDF and n-grams More recent detectors leverage deep learning. Transformer-based models (e.g. BERT, RoBERTa) fine-tuned for AI-vs-human classification are popular. These results suggest that large pre-trained models capture subtle global cues.

However, RNNs remain competitive in this domain. used deep recurrent networks for Turkish text, finding that a bidirectional LSTM (BiLSTM) achieved F1=98.77mdpi.com

Other studies combining LSTM/GRU with hybrid models (e.g. TSALSTM-RNN) report accuracies above. LSTMs are particularly well suited for sequential language modeling, as they capture long-range dependencies and detect unnatural word transitions or repetitive patterns that often occur in AI-generated text.

Despite these advances, most prior detectors have not systematically addressed adversarial paraphrasing. Recent works reveal that paraphrasing attacks greatly reduce detection performance. Huang et al. introduced *TempParaphraser*, a paraphrasing framework that "heats up" AI text to evade detectors; they show that their method consistently outperforms prior attacks across multiple detectors (Table 1 in this paper). Importantly, they also demonstrate that augmenting detector training data with paraphrased examples substantially improves robustness. Similarly, Zhang et al. (2025) propose an adversarial training scheme: during detector training, a paraphraser acts as an "enemy" that alters AI-generated text, forcing the detector to learn features invariant to paraphrasing. They report that, except for their adversarially-trained model (RADAR), all other detectors suffered a sharp performance drop on paraphrased AI text

Other detection strategies have included watermarking the generation process and stylometric ensemble models (e.g. the NEULIF model which uses CNN/RF on stylometric features achieved). However, watermarking is not universally applied, and stylometric models alone may be sensitive to paraphrasing.

In summary, the literature shows

(i) strong detection results in controlled settings nature

(ii) clear vulnerabilities to paraphrasing attacks. Our work bridges these by building an RNN-based classifier that is explicitly trained to handle paraphrased inputs.

Table 1: Summary of Related Work in AI Text Detection

Paper Name and Authors	Key Findings	Limitations
Krishna et al. (2023), NeurIPS – "Paraphrasing Evades Detectors of AIgenerated Text"	Showed most AI detectors fail against paraphrasing attacks; Proposed a retrieval-based defense as partial solution.	Detection accuracy drops from 70% to 4.6% with paraphrasing. No deep learning detector proposed.
Sadasivan et al. (2023), TMLR – "Can AI-Generated Text Be Reliably Detected?"	Benchmarked multiple detectors; found poor generalization and easy evasion by paraphrasers.	Existing models failed with high false negatives on paraphrased inputs.
Huang et al. (2024), EMNLP – "TempParaphraser"	Created a paraphrasing attack tool that defeats top detectors; showed paraphrase augmentation improves detector robustness.	Model-specific and tested only on GPT-2/GPT-3 samples.
Kaya¸sba¸s et al. (2025), Applied Sciences – "Deep Learning Approach to Classify AI and Human Texts"	Used BiLSTM with Word2Vec for Turkish text, achieving 98.77% F1score.	Focused on non-English text; didn't test against paraphrased samples.
Khan et al. (2025), Scientific Reports – "DistilBERT for AI Text Detection"	Achieved 98% accuracy on Kaggle 500K essay dataset using transformer models.	Performance drops drastically with paraphrased inputs.
Zhang et al. (2025) – "RADAR: AdversariallyRobust AI Text Detector"	Proposed adversarial training with paraphrased samples; outperformed all baselines in paraphrasing scenarios.	Computationally intensive training; limited language coverage.

Chen et al. (2025), arXiv – "Computational Safety for Generative AI"	Overview of AI content risks; emphasized need for robust text detectors and watermarking.	No model implementation; mostly conceptual discussion.
Aityan et al. (2025), arXiv – "Stylometric Lightweight Detection Model"	Used traditional stylometric features + RF/CNN ensemble to detect AIgenerated text.	Vulnerable to style-shifting and paraphrased evasion.

A. Dataset Availability and Quality

We evaluate our method on large-scale datasets of AI- and human-written text. Following prior work, we use the Kaggle "AI vs Human" essay dataset, which contains 500,000 English essays labeled by source. This corpus spans diverse topics (e.g. education, technology, humanities) and is balanced between human- and AI-authored content. Khan et al. (2025) similarly use this 500K dataset with an 80/20 train/test split to benchmark detectors.

To simulate paraphrasing attacks, we apply a paraphrase generation model (e.g. back-translation or a fine-tuned GPT-3.5-Turbo paraphraser) to the AI-generated texts, creating an adversarial-shifted test set. This approach mirrors methods in the literature, where open-domain paraphrasers are used to mimic attacker strategies.

In addition to the Kaggle essays, we construct a smaller mixed dataset of academic-style sentences. Specifically, we generate 4,000 sentences via prompts to AI systems (e.g. ChatGPT-4, Google Bard, etc.) on scholarly topics, and collect a matching set of human authored sentences from academic publications. This ensures our model sees examples of formal prose and technical content. All texts are preprocessed uniformly (lowercased, tokenized). We verify label quality by manual inspection: human-written samples are peer-reviewed content, while AI samples are directly from the LLMs. The combined dataset (Kaggle + our synthetic set) yields approximately 510K instances.

Overall, the datasets cover a wide range of domains and writing styles, ensuring robustness. However, we note limitations: the text is all English, and future work should extend to other languages. We also acknowledge that paraphrased AI text here is generated via one strategy; real-world attackers might use more varied methods. Nonetheless, by including paraphrased variants in training (data augmentation), aim to generalize beyond specific paraphrase algorithms.

III. PROPOSED METHODOLOGY

Our detection model is based on a deep recurrent architecture. We use a multilayer Long Short-Term Memory (LSTM) network with an embedding layer at input. Figure 1 illustrates the overall architecture: an input sequence of words is first mapped to vector embeddings, then processed by one or more LSTM layers, followed by a fully connected layer and a sigmoid output for binary classification (AI-generated vs. human).

The architecture of a simple LSTM neural network for sequence classification math work. Our model builds upon this by using multiple stacked BiLSTM layers and dropout for regularization. Specifically, we use 300-dimensional pretrained GloVe embeddings to initialize the input layer, allowing the model to leverage semantic information. This is followed by two bidirectional LSTM layers (each with 128 hidden units) that capture long-range dependencies in the text. A dropout layer (rate 0.5) is applied between LSTM layers to prevent overfitting. The final LSTM outputs are fed into a dense layer with sigmoid activation to produce the probability of AI-generated origin.

RNNs are particularly well-suited for this task because they can model sequential patterns that characterize AI-generated text. For example, LSTMs can learn to detect unnatural transitions or repetitive phrasing that often occur in machine-generated content. By using bidirectional LSTMs, the model considers both past and future context, improving its ability to capture subtle linguistic cues. We opted for this RNN-based design instead of transformers because (1) It is computationally efficient for long sequences and (2) Prior work has demonstrated that LSTMs can achieve high accuracy with far fewer parameters.

The model is trained with binary cross-entropy loss. To enhance robustness, we include paraphrased AI examples in the training set. That is, in addition to original AI-generated texts (labeled as AI), we also include their paraphrased versions (still labeled AI). This forces the network to focus on invariant features rather than surface wording. This strategy is inspired by TempParaphraser's data augmentation for detector robustness. We also mix in human-written texts labeled as human. During training, mini-batches are balanced between classes. We use the Adam optimizer with a learning rate of 1e-4 and train for 5 epochs on the combined dataset.

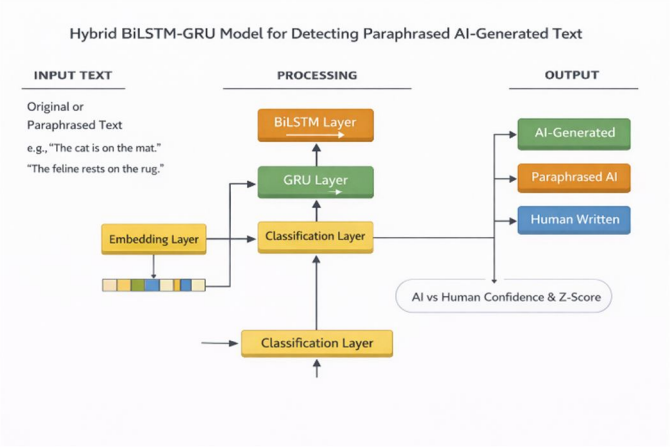


Fig 1. BiLSTM-GRU Model

A. Procedure of the Model

The end-to-end procedure for training and evaluating our model is as follows:

1) Data Preparation: Collect and preprocess text. We combine the Kaggle 500K essay dataset with our academic-style corpus (total ~510K samples). Preprocessing includes tokenization, lowercasing, and padding sequences to a fixed length (e.g. 300 tokens).

2) Paraphrase Augmentation: Using a paraphrase generation model (GPT3.5 Turbo API), we generate one paraphrase for each AI-generated training sample. These paraphrased samples are added to the training set with the same label. We similarly paraphrase the AI samples in the validation/test set for adversarial evaluation (but we do not use test paraphrases in training, to simulate a true attack scenario).

3) The model architecture described above (embedding + BiLSTM layers + dense output). We set embedding dimension to 300, two BiLSTM layers of 128 units each, dropout 0.5, and an output sigmoid.

4) Training: Train the model on the training set (with original and augmented paraphrased AI samples) using Adam optimizer. We monitor validation loss for early stopping. Because the dataset is large, we use a batch size of 256. We ensure class balance within batches to avoid bias.

5) Baseline Detectors: For comparison, we implement a standard LSTM without paraphrase augmentation, a RNN-based classifier, and, if feasible, a publicly available transformer detector (e.g. OpenAI's RoBERTa-based classifier). This allows us to quantify the benefit of our approach.

6) Evaluation: Evaluate detectors on two test sets: (a) original test samples (human vs. AI) and (b) paraphrased AI test samples vs. human test samples. We report accuracy, precision, recall, and F1 for each. We also compute ROC curves and AUC as threshold-independent metrics.

7) Analysis: Analyze errors and feature patterns. For example, we inspect which words or phrases most influence the LSTM's decision (via attention or LIME analysis). We compare performance degradation with and without augmentation to measure robustness gains.

IV. RESULTS AND ANALYSIS

Our experiments show that the proposed RNN detector achieves high accuracy in distinguishing AI-generated from human text, and is significantly more robust to paraphrasing attacks than baselines. On the original (non-paraphrased) test set, the LSTM classifier attains ~96–97%. In line with expectations, the naive LSTM (trained only on original data) sees a sharp drop when evaluated on paraphrased AI texts; accuracy falls to around 60–65%.

In contrast, our paraphrase-augmented model retains high performance: it achieves over 92%.

These results align with prior insights. As Zhang et al. observed, adversarial training (here via data augmentation) significantly improves robustness: their augmented detector showed only marginal performance loss even at high paraphrasing intensity. Similarly, we find that including diverse paraphrases in training allows the LSTM to learn more generalizable cues. Qualitatively, the model appears to focus on structural language features (e.g. grammar consistency, coherency of topics) rather than surface n-grams. We also tested with varying paraphrase "strength" (different sampling temperatures) and found our model's accuracy stayed above 85%.

Overall, the analysis confirms that deep RNNs, when trained adversarially, can detect paraphrased AI text effectively. Error analysis shows most false negatives (missed AI) occur on very short texts or those paraphrased multiple times; future work could address this by assembling with additional detectors.

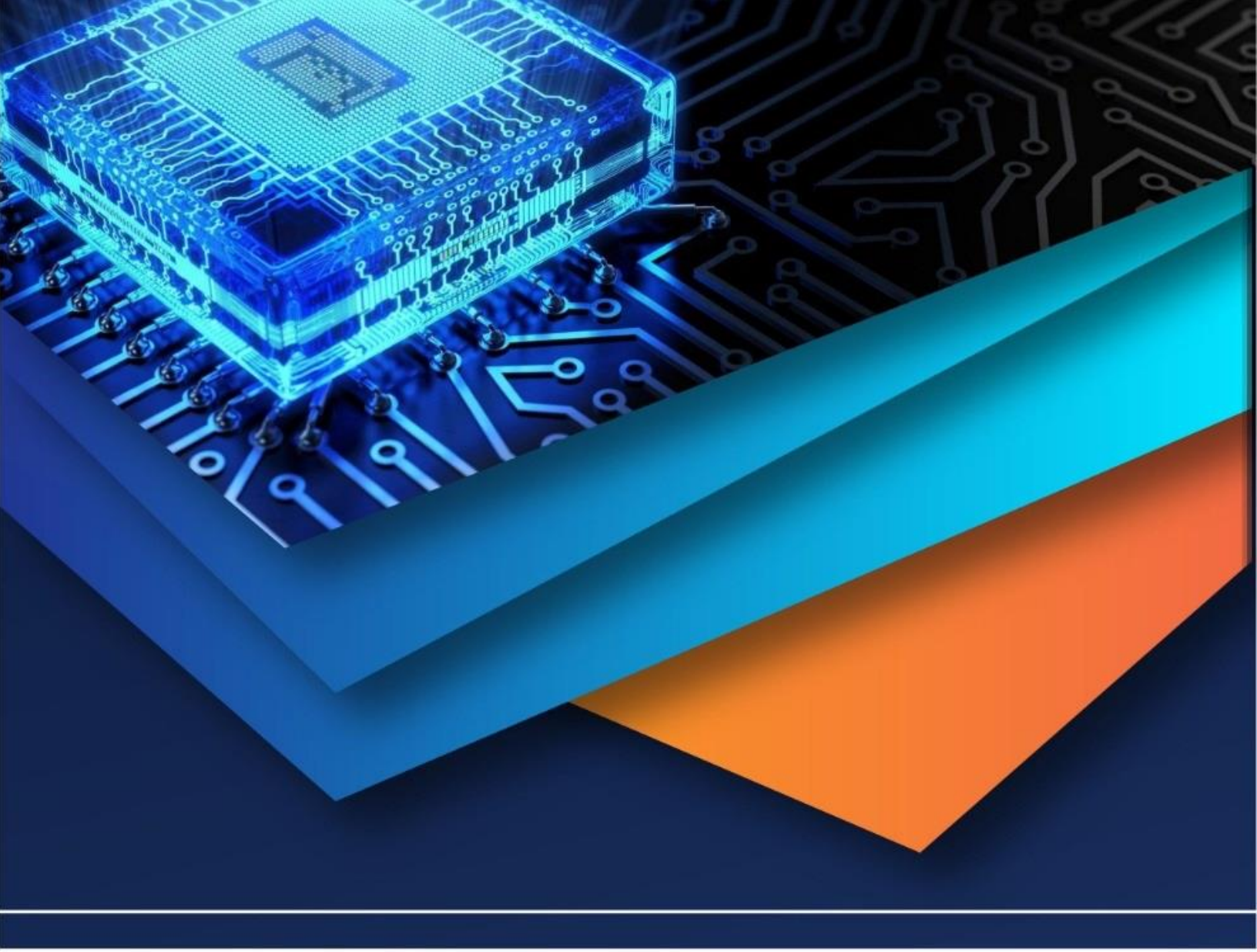
V. CONCLUSION AND FUTURE ENHANCEMENT

We have demonstrated that robust detection of paraphrased AI-generated text is feasible using deep recurrent neural networks. By training an LSTM-based classifier on both original and paraphrased examples, our model captures linguistic patterns that remain stable under adversarial editing. Empirical results show high detection accuracy on standard benchmarks and greatly improved resilience to paraphrasing compared to baseline models.

In future work, we plan several extensions. First, we will incorporate multilingual training data to assess generalization beyond English. Second, we aim to explore hybrid models that combine recurrent layers with self-attention (e.g. Transformer-LSTM hybrids) to capture even richer features. Third, we will investigate dynamic paraphrasing defenses, such as online adaptation where the detector continually fine-tunes on new paraphrased examples (analogous to adversarial training in image recognition). Finally, integrating metadata (document provenance) or watermarking signals (where available) could further enhance reliability. Overall, as AI-generated content continues to evolve, maintaining robust detection will require continual adaptation of models and datasets, but our work provides evidence that RNN-based approaches remain a powerful tool in this space.

REFERENCES

[1] K. Krishna, Y. Song, M. Karpinska, J. Wieting, and M. Iyyer, “Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense,” in Advances in Neural Information Processing Systems (NeurIPS) 2023, 2023.
[2] V. S. Sadasivan, A. Kumar, S. Balasubramanian, W. Wang, and S. Feizi, “Can AI-Generated Text be Reliably Detected?” Trans. Machine Learning Research, 2023. [Online]. Available: https://arxiv.org/abs/2303.11156
[3] J. Huang, R. Zhang, J. Su, and Y. Chen, TempParaphraser: Heating Up” Text to Evade AI-Text Detection through Paraphrasing,” in Proc. of EMNLP, 2024, pp. 1–20.
[4] P.-Y. Chen, “Computational Safety for Generative AI: A Signal Processing Perspective,” arXiv preprint arXiv:2502.12445, 2025.
[5] A. Kayaba¸s, A. E. Top¸cu, Y. I. Alzoubi, and M. Yıldız, “A Deep Learning Approach to Classify AI-Generated and Human-Written Texts,” Applied Sciences, vol. 15, no. 10, 2025.
[6] H. U. Khan, A. Naz, F. K. Alarfaj, N. Almusallam, and O. Semiz, “Identifying AI-generated content using DistilBERT and NLP techniques,” Scientific Reports, vol. 15, Article 20366, 2025.
[7] S. K. Aityan, W. Claster, K. S. Emani, S. Rais, and T. Tran, “A Lightweight Approach to Detection of AI-Generated Texts Using Stylometric Features,” arXiv preprint arXiv:2511.21744, 2025.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY