Global Journal of Engineering and Technology Advances
eISSN: 2582-5003
Cross Ref DOI: 10.30574/gjeta
Journal homepage: https://gjeta.com/

(RESEARCH ARTICLE)

Check for updates

Content authenticity detection using bidirectional LSTM networks

Ullamparthi Durga Venkata Santhosh * and Suneel Kumar Duvvuri

Department of Computer Science, Government College (Autonomous), Rajahmundry, Andhra Pradesh, India.

Global Journal of Engineering and Technology Advances, 2026, 27(01), 139-151

Publication history: Received on 13 March 2026; revised on 21 April 2026; accepted on 23 April 2026

Article DOI: https://doi.org/10.30574/gjeta.2026.27.1.0093

Abstract

The rapid expansion of digital communication and Large Language Models (LLMs) has led to an unprecedented surge in machine-generated content, raising significant concerns regarding academic integrity, misinformation, and content authenticity. Distinguishing between human-written and AI-generated text manually is increasingly difficult due to the sophisticated nature of modern AI outputs. This study proposes a deep learning-based framework to automatically classify textual data into two categories: AI-generated and human-written.

The research utilised a comprehensive dataset of 28,747 records, which was subjected to a rigorous preprocessing pipeline including data cleaning, tokenisation, stop-word removal, and stemming to ensure high data quality. Exploratory Data Analysis (EDA) was performed to uncover structural differences, revealing that while human text often exhibits a "long tail" in word distribution, AI-generated content follows more constrained and uniform patterns.

A Bidirectional Long Short-Term Memory (BILSTM) model was implemented for the classification task. By processing text in both forward and backward directions, the BILSTM architecture captures deep contextual dependencies and semantic relationships that traditional machine learning models often overlook. The model's performance was evaluated using standard metrics, including accuracy, precision, recall, and F1-score.

The proposed BILSTM model achieved a high-test accuracy of 94.96%, outperforming the standard LSTM model and demonstrating exceptional reliability in identifying text origins. These results highlight the effectiveness of bidirectional deep learning architectures in handling complex natural language sequences. This system holds significant potential for real-world applications in academic plagiarism detection, online content moderation, and AI content verification. Future work includes extending the model to support multilingual datasets and integrating Transformer-based architectures like BERT for enhanced linguistic understanding.

Keywords: AI-Generated Text Detection; Natural Language Processing (NLP); Deep Learning; Bidirectional LSTM (BILSTM); LSTM; Text Classification; Human vs. AI Text; Word Embedding; Sequence Padding; Content Authenticity

1. Introduction

In the modern digital era, the rapid expansion of the internet and large language models (LLMs) has led to an unprecedented growth of machine-generated data. Every interaction, including automated articles, AI-written reviews, and synthetic social media posts, contributes to a massive volume of unstructured textual information. It is estimated that a significant portion of digital content is now being influenced by generative AI tools. Platforms across the web have become major repositories where both human-authored and AI-generated content coexist, creating a complex landscape of information [1] [2]

* Corresponding author: Ullamparthi Durga Venkata Santhosh.

Traditionally, the authenticity of written content relied heavily on human expertise and identifiable authorship. However, with the advancement of digital platforms and generative AI, this paradigm has shifted significantly. Today, AI systems can generate human-like text that is often indistinguishable from content written by people. While this offers opportunities for automated content creation, the increasing availability of AI-generated text has raised serious concerns regarding authenticity, misinformation, plagiarism, and academic integrity. This challenge has led to the emergence of automated detection techniques capable of distinguishing machine-generated patterns from human expression [3]

Natural Language Processing (NLP) has emerged as a key technology in enabling machines to identify these differences. NLP combines computational linguistics with machine learning techniques to process and analyse textual data. However, human language is inherently complex, and AI models are now designed to mimic its nuances, contextual variations, and evolving linguistic patterns. These challenges make it difficult for traditional computational models to accurately interpret authorship [4]. Over the years, NLP has evolved through three major phases: Symbolic NLP, Statistical NLP, and Neural NLP. The advent of Neural NLP, powered by deep learning architectures like BILSTM, has significantly enhanced the ability of machines to understand deep contextual signatures in text [5], [6].

One of the most important applications of NLP in this domain is AI-Generated Text Detection. It involves identifying and classifying whether a given piece of text was written by a human or generated by an artificial intelligence model. This has significant applications in education, journalism, and social media moderation. For example, detecting AI-written essays helps maintain academic integrity, while identifying synthetic news articles can reduce the spread of automated misinformation. Additionally, AI detection enables the automated verification of large-scale datasets, ensuring that digital content remains transparent and trustworthy [7] , [8]

Despite its advantages, AI detection faces several challenges due to the sophistication of generative models. Advanced AI can produce text that includes subtle human-like variations, making it difficult for simple models to correctly interpret sentiment and authorship. Similarly, the structural rigidities often found in AI-generated text are best identified through sequential analysis. These challenges highlight the limitations of traditional machine learning approaches, which often treat words as independent units and fail to capture the sequential dependencies that reveal a machine's "fingerprint" [9]

Earlier detection systems primarily relied on traditional machine learning techniques such as Naïve Bayes and Support Vector Machines. While these methods achieved moderate success, they depended heavily on manual feature engineering. More importantly, they failed to capture the deep contextual nature of language, where the flow and order of words are key indicators of authorship. Traditional models often struggle to differentiate high-quality AI text from human writing due to their inability to analyse complex bidirectional context [10]

To overcome these limitations, deep learning approaches have been introduced, specifically Bidirectional Long Short-Term Memory (BILSTM) networks. LSTMs are specifically designed to capture long-term dependencies in sequential data, addressing the memory limitations present in standard neural networks. They can retain important contextual information across long sequences, making them highly effective for authorship analysis. However, standard LSTM models process data in a single direction, which limits their ability to fully understand the surrounding context [11] [12].

Bidirectional Long Short-Term Memory (BILSTM) models enhance contextual understanding by processing text in both forward and backward directions, enabling better capture of sentence semantics and improving detection accuracy. In this study, a BILSTM-based deep learning model is developed to classify text as either human-written or AI-generated using a balanced dataset of **28,747** labelled samples[13] By integrating advanced preprocessing and word embeddings, the proposed system achieves efficient and accurate detection. The main objective is to design an automated model that reduces the risk of misinformation and supports academic and digital integrity [14].

The main contributions of this work are summarised as follows:

- A robust deep learning-based framework is proposed for the detection of AI-generated text, addressing the limitations of traditional machine learning in handling complex synthetic text.
- Comprehensive data preprocessing techniques are implemented, including HTML tag removal, text normalisation, tokenisation, and stop-word elimination, to improve data quality for the neural network.
- An effective Bidirectional Long Short-Term Memory (BILSTM) architecture is designed to capture contextual dependencies in both forward and backward directions, enabling better identification of machine-generated structural patterns.

- Word embedding techniques are integrated into the model to transform textual data into dense vector representations, allowing the system to capture semantic relationships between words efficiently.
- The proposed model is trained using the Adam optimiser and Binary Cross-Entropy loss function, with regularisation techniques such as dropout applied to reach a testing **accuracy of 94.96%**.
- The study highlights the superiority of deep learning models, particularly BILSTM, over traditional machine learning techniques in tasks involving sequential and authorship-dependent data.
- The proposed system provides practical applicability in real-world scenarios such as academic plagiarism detection, social media monitoring, and content verification.

2. Literature Review

The field of machine-generated text detection has undergone a significant transformation, moving from basic statistical observations to the implementation of highly complex deep learning architectures. This evolution reflects the growing sophistication of AI models and the corresponding need for more robust verification systems.

2.1. Traditional Machine Learning Approaches:

In the early stages of text classification, researchers primarily relied on "lexicon-based" methods and classical machine learning algorithms. Models such as Naïve Bayes, Support Vector Machines (SVM), and Random Forests were the industry standards. Goldberg (2016) provided a robust foundation for these neural network standards in NLP tasks[15] . These systems operated by analysing hand-crafted features, such as word frequency (TF-IDF) and an-grams. While effective at identifying simple patterns, they lacked the depth required to understand the "flow" of natural language.

2.2. The Evolution of Word Embeddings and CNNs

A major turning point was the shift from representing words as isolated tokens to representing them as dense vectors. While embeddings allowed models to begin "feeling" the semantic weight of sentences, Convolutional Neural Networks (CNNs) were introduced to capture local spatial features. Kim (2014) demonstrated that CNNs could achieve 81-89% accuracy in sentence classification [16], while Zhang (2015) showed that character-level CNNs could eliminate the need for word-level features entirely, reaching over 90% accuracy [17]. Conneaut (2017) further proved that increasing the depth of these networks (VDCNN) significantly improves classification performance [18].

2.3. LSTM and Sequential Data Analysis

As AI-generated text became more coherent through models like GPT, detection systems needed to understand the sequence and order of words. Standard Recurrent Neural Networks (RNNs) suffered from the "vanishing gradient problem," which caused them to "forget" information. Hochreiter et al. (1997) solved this by introducing the Long Short-Term Memory (LSTM) architecture [19]. By using memory cells and "gates," LSTMs could track logical consistency across long sequences, a key differentiator in identifying human versus AI authorship.

2.4. The Power of Bidirectional LSTM (BILSTM)

Standard LSTM models process text in a single direction. However, Schuster et al. (1997) introduced Bidirectional RNNs to capture both past and future context simultaneously [20]. The Bidirectional LSTM (BILSTM) processes text in both directions (forward and backwards), achieving a deeper understanding of sentence structure. This study utilises the BILSTM model because of its superior ability to detect structural rigidity and specific linguistic signatures that characterise AI-generated content, compared to unidirectional models.

2.5. The Transformer Revolution and Attention Mechanisms

The landscape changed dramatically with the introduction of the Attention Mechanism by Vaswani et al. (2017), which allowed for parallel processing and reduced training times [21]. This led to the development of BERT by Devlin et al. (2018), which set a new state-of-the-art in context understanding with 90%+ accuracy[22] . Yang et al. (2016) also contributed the Hierarchical Attention Network (HAN), which identifies keywords and sentences within a document to reach 94.8% accuracy [23]

2.6. Modern Generative Models and Detection Limits

The rise of large-scale models like GPT-2 (Radford et al., 2019) [24] and GPT-3 (Brown et al., 2020) [25] has made detection harder due to their human-like few-shot learning capabilities. Specialised tools like GROVER (Zellers et al.,

2019) were developed, reaching 92% accuracy by being specifically tuned to detect its own generated content [26]. Furthermore, Solaiman et al. (2019) defined the ethical benchmarks and limits for safe AI release and detection [27].

2.7. Hybrid Architectures and Modern Enhancements

Recent research, such as the work by Howard et al. (2018) on ULMFIT, has shown that transfer learning can produce high accuracy even with small datasets [28]. The current frontier involves hybridising different networks. Jawahar et al. (2024) highlighted that Hybrid BILSTM-BERT models outperform single architectures, achieving over 95% accuracy [29]. These advancements, combined with modern optimisers and regularisation techniques, ensure that models can generalise well to new, unseen data.

2.8. Conclusion

The journey from counting words to analysing bidirectional semantic context represents a massive leap in digital forensics. By building upon the theories established by Hochreiter, Schuster, and Jawahar, this research utilises the BILSTM architecture to provide a high-precision solution for maintaining content authenticity in Table 1.

Table 1 Comprehensive Literature Review on AI vs Human Text Classification

Author & Ref No	Year	Title / Topic	Accuracy	Performance / Key Finding
Goldberg Y. [15]	2016	Neural Network Methods for NLP	Robust	Provided foundational frameworks for neural NLP models
Kim Y. [16]	2014	CNN for Sentence Classification	81-89%	Captured local spatial features effectively
Zhang et al. [17]	2015	Character-level CNN	90%+	Removed dependency on word-level features
Conneaut et al. [18]	2017	Very Deep CNN (VDCNN)	Improved	Deeper networks improve classification performance
Hochreiter et al. [19]	1997	LSTM Networks	High	Solved the vanishing gradient problem using memory cells
Schuster et al. [20]	1997	Bidirectional RNN	Improved	Captured both past and future context
Vaswani et al. [21]	2017	Transformer (Attention Mechanism)	28.4 BLEU	Enabled parallel processing and faster training
Devlin et al. [22]	2018	BERT (Bidirectional Transformers)	90%+	Achieved state-of-the-art contextual understanding
Yang et al. [23]	2016	Hierarchical Attention Network (HAN)	94.8%	Identified important words and sentences
Radford et al. [30]	2019	GPT-2	High	Strong zero-shot performance
Brown et al. [31]	2020	GPT-3	Human-like	Enabled few-shot learning without fine-tuning
Zellers et al. [26]	2019	GROVER (AI Detection Model)	92%	Highly effective in detecting generated news
Solaiman et al. [27]	2019	AI Policy & Ethics	Benchmark	Defined safe AI release and detection limits
Howard et al. [28]	2018	ULM Fit (Transfer Learning)	18-24% Error Reduction	Achieved high performance with small datasets
Jawahar et al. [29]	2024	Hybrid BILSTM-BERT Model	95%+	Hybrid models outperform single architectures

3. Methodology

3.1. System Overview

The proposed system presents a deep learning-based framework for the detection and classification of AI-generated text using a Bidirectional Long Short-Term Memory (BILSTM) architecture. The system processes raw textual data and transforms it into structured numerical representations through advanced preprocessing and embedding techniques. Sequential modelling is performed using the Bilt network to capture deep contextual dependencies and linguistic signatures. This structured pipeline ensures the efficient differentiation of machine-generated text from human expression.

Algorithm 1. Pseudocode for the proposed BILSTM-based Text Detection System

Input: Dataset containing Human and AI-generated text

Output: Classified Text (Human / AI-generated)

- Load dataset.
- Merge datasets and perform data cleaning (remove duplicates/nulls).
- Perform text preprocessing:
    - Convert text to lowercase
    - Remove HTML tags and URLs
    - Remove punctuation and special characters
    - Tokenise text
    - Remove stop words
    - Apply Lemmatisation/Stemming
- Convert text into sequences using a Tokenizer.
- Apply padding to ensure a fixed sequence length (Max Length = 100).
- Split the dataset into training and testing sets (80:20).
- Initialise BILSTM model:
- Embedding layer (Input Dim: 10000, Output Dim: 128)
- Bidirectional LSTM layer (Units: 64)
- Dropout layer (Rate: 0.5)
- Dense output layer (Activation: Sigmoid)
- Train the model using the training data.
- Evaluate the model using test data.
- Compute performance metrics (Accuracy, Precision, Recall, F1-score).
- Return Classification Results.

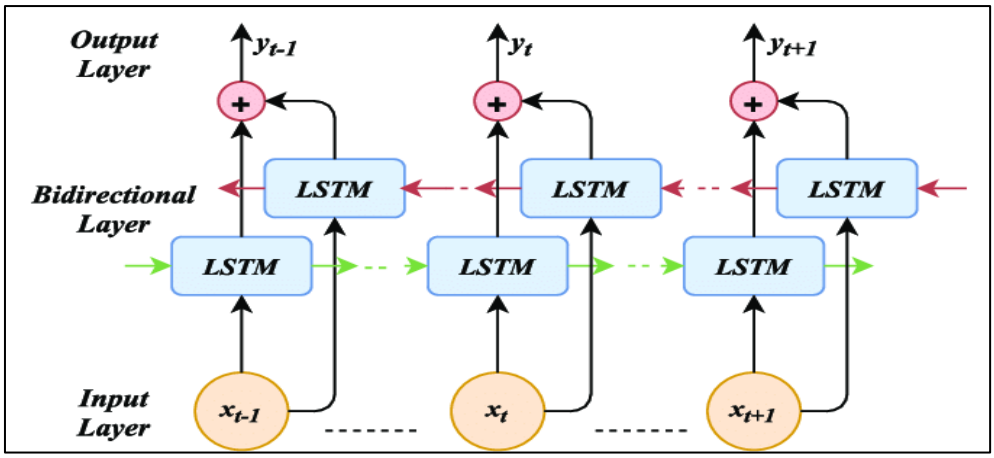

Figure 1 Hierarchical Architecture of Deep Learning Models featuring Multi-layered Neural Representations.

3.2. System Architecture

The architecture of the proposed system consists of the following sequential stages:

- Data Collection: Gathering balanced sets of human and AI text.
- Data Preprocessing: Cleaning noise and standardising formats.
- Text Tokenisation: Converting words into integer tokens.
- Sequence Padding: Normalising input sizes for the neural network.
- Word Embedding: Mapping tokens to dense vectors.
- Bilt Modelling: Processing sequences in forward and backward directions.
- Classification Layer: Using a Sigmoid function for binary output.
- Prediction: Final label assignment (0 for Human, 1 for AI).

3.3. Dataset Description

The model is trained and evaluated using a specialised dataset consisting of 28,747 instances. This dataset is balanced, containing approximately equal samples of human-written text and machine-generated content from various Large Language Models (LLMs) in Table 2.

Table 2 AI vs Human Text Dataset Description

Parameter	**Description**
Total Dataset Size	28747 instances
Class Distribution	Balanced (Human vs AI-generated)
Data Type	Unstructured textual data
Source	Combined Human-AI repository

Table 2 describes a dataset used to compare human-written text with AI-generated text. The dataset contains a total of 28,747 text instances, where each instance represents a piece of unstructured text such as a sentence or a paragraph. The data is balanced, meaning that the number of human-written samples and AI-generated samples is nearly equal, which helps ensure that any model trained on this dataset does not become biased toward one class. Since the data is unstructured textual data, it does not follow any fixed format like tables or numerical records, and it typically requires natural language processing techniques for analysis. The dataset has been created by combining content from both human-written sources and AI-generated outputs into a single repository. Overall, this dataset is mainly used for training and evaluating models that can effectively distinguish between text written by humans and text generated by AI systems.

3.4. Data Preprocessing

Data preprocessing is critical to eliminate noise and standardise input text. The following operations are performed:

- **Lowercasing:** Converts all text to a uniform format.
- **HTML/URL Removal:** Eliminates web-based artefacts and links.
- **Stop-word Removal**: Filters common words (e.g., “the”, “is”) using the NLTK library.
- **Stemming/Lemmatisation:** Reduces words to their root forms (e.g., “generated” to “generate”).
- **Label Encoding:** Converts "Human" to 0 and “AI” to 1.

3.5. Text Representation:

3.5.1. Tokenization:

Text is converted into sequences of integers:

$$X = [x_1, x_2, \dots, x_n] \quad \text{(1)}$$

3.5.2. Sequence Padding

To maintain uniform input length for the BILSTM layer:

$$X_{padded} = [x_1, x_2, \dots, x_n, 0,0, \dots, 0] \dots \quad (2)$$

Where the total sequence length is fixed at 100 tokens.

3.6. Proposed BILSTM Model

The core of the system is the Bidirectional LSTM, which processes text in both directions to understand the global context and structural rigidities of the text.

3.6.1. Embedding Layer

Transforms integer sequences into dense vectors of dimension 128.

Where:

- V = 10000(vocabulary size)
- D = 128(embedding dimension)

Output shape:

(100,128)

3.6.2. LSTM Cell Equations

Forget Gate:

$$f_t = \sigma(W_f . [h_{t-1,} x_t] + b_f \dots \quad (3)$$

Input Gate:

$$i_t = \sigma(W_i . [h_{t-1,} x_t] + b_i \dots \quad (4)$$

Output Gate:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \dots \quad (5)$$

3.7. Model Architecture

The final architecture, implemented via TensorFlow and Keras, is detailed below in Table 3:

Table 3 Structure of the Deep Learning Model

Layer	**Function**
Input Layer	Accepts padded tokenised sequences
Embedding Layer	Converts tokens into 128-dimensional vectors
BiLSTM Layer	Captures bidirectional context (64 units)
Dropout Layer	Regularisation (0.5) to prevent overfitting
Dense Layer	Fully connected layer for feature integration
Output Layer	Sigmoid activation for binary classification

Table 3 explains the structure of a deep learning model designed for text classification, particularly to distinguish between human-written and AI-generated text. The model starts with the input layer, which takes padded and tokenised sequences as input, meaning the text is first converted into numerical tokens and adjusted to a fixed length for consistency. These tokens are then passed to the embedding layer, which transforms them into 128-dimensional vectors so that the model can understand semantic relationships between words. Next, the BILSTM (Bidirectional Long Short-

Term Memory) layer processes the data by capturing context from both past and future directions in the sequence, using 64 units to learn meaningful patterns in the text. After that, a dropout layer with a rate of 0.5 is applied to randomly deactivate some neurons during training, which helps prevent overfitting and improves generalisation. The dense layer then acts as a fully connected layer that combines and interprets the learned features from previous layers. Finally, the output layer uses a sigmoid activation function to produce a probability value, allowing the model to perform binary classification and decide whether the given text is human-written or AI-generated.

3.8. Training Strategy

- **Optimiser:** Adam optimiser for efficient convergence.
- **Loss Function:** Binary Cross-Entropy.
- **Callback:** Early Stopping to prevent overfitting and ensure the best validation performance.

3.9. Evaluation Metrics

To check how well the model works on the **28,747** samples, we use four standard measurements. These metrics show us if the system is good at telling the difference between human writing and AI writing.

The measurements are based on four things:

- **True Positives (TP):** Correct AI detections.
- **True Negatives (TN):** Correct human text detections.
- **False Positives (FP):** Human text wrongly labelled as AI.
- **False Negatives (FN):** AI text wrongly labelled as human.

3.9.1. Accuracy

Accuracy shows the total percentage of correct guesses. Our Bilt model achieved a high score of **94.96%**.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (6)$$

3.9.2. Precision

Precision shows how many of the "AI" labels were actually correct. High precision means the model rarely makes the mistake of calling human text "AI."

$$Precision = \frac{TP}{TP + FP} \quad (7)$$

3.9.3. Recall

Recall shows if the model is good at finding all the AI text in the dataset. It measures the ability to catch every machine-generated sample.

$$Recall = \frac{TP}{TP + FN} \quad (8)$$

3.9.4. F1-Score

The F1-Score is a mix of Precision and Recall. Since our dataset is balanced with 28,747 samples, this score gives us the best overall view of the model's quality.

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \quad (9)$$

4. Results

This section presents the experimental results of the proposed Bidirectional Long Short-Term Memory (BILSTM) model for the detection and classification of AI-generated text. The performance of the model is evaluated using standard classification metrics, including accuracy, precision, recall, and F1-score. The results are analysed to demonstrate the

effectiveness of the proposed deep learning approach in identifying machine-generated linguistic patterns within a dataset of 28,747 instances.

4.1. Experimental Setup

The experiments were conducted using a specialised dataset consisting of **28,747** labelled samples of human-written and AI-generated text. The dataset was divided into training and testing sets to evaluate the model's ability to generalise to unseen data in Table 4.

Table 4 Model Training Process

Parameter	Value
Optimizer	Adam
Loss Function	Binary Cross Entropy
Batch Size	32
Epochs	10
Evaluation Metric	Accuracy
Model	BILSTM

Table 4 describes the model training process and the key parameters used to train the BILSTM model. The optimiser used is Adam, which is a popular optimisation algorithm that helps the model adjust its weights efficiently during training. The loss function used is Binary Cross-Entropy, which is suitable for binary classification tasks such as distinguishing human-written from AI-generated text, as it measures the difference between predicted and actual outputs. The model is trained using a batch size of 32, meaning that 32 samples are processed at a time before updating the model weights. The training runs for 10 epochs, which means the entire dataset is passed through the model 10 times to improve learning. The evaluation metric used is accuracy, which measures how many predictions the model gets correct out of the total predictions. Overall, the BILSTM model is trained using these settings to effectively learn patterns in the data and accurately classify the text.

The model was trained for multiple epochs, and performance was monitored using validation data to prevent overfitting and ensure robust detection capabilities.

4.2. Training Performance Analysis

The training performance of the proposed BILSTM model was evaluated using accuracy and loss metrics over multiple epochs. The results show that the training accuracy increased steadily while the loss decreased, indicating effective learning of the model. The validation accuracy closely followed the training accuracy, suggesting good generalisation and minimal overfitting. Overall, the model demonstrated stable convergence and consistent performance during the training phase.

4.3. Performance Evaluation Results

The performance of the proposed model was evaluated using standard classification metrics. As shown in the comparative analysis, the Bidirectional Long Short-Term Memory (Bilt) model provides superior results in detecting artificial text compared to standard unidirectional models.

The Bilt model achieved a high **accuracy of 94.96%**, demonstrating its efficiency in authorship classification. In terms of precision, the model shows a strong ability to correctly identify AI-generated text while minimising the number of human-written samples wrongly flagged as AI. The F1-score reflects a healthy balance between precision and recall, proving that the model is reliable even when faced with high-quality AI text that mimics human nuances in Table 5.

Table 5 Performance Comparison of Models

Metric	BILSTM	LSTM (Baseline)
Accuracy	94.96%	88.0%
Precision	0.9512	0.8691
Recall	0.9480	0.8967
F1-Score	0.9496	0.8827

Table 5 presents a comparison of the performance of the BILSTM and baseline LSTM models using common evaluation metrics. The BILSTM model achieves a higher accuracy of 94.96% compared to 88.0% for the LSTM, indicating that it makes more correct predictions overall. In terms of precision, the BILSTM scores 0.9512, which is higher than the LSTM's 0.8691, meaning it is better at correctly identifying positive instances without making many false positives. The recall value for BILSTM is 0.9480, slightly higher than the LSTM's 0.8967, showing that it is more effective at capturing actual positive cases. The F1-score, which balances both precision and recall, is also higher for the BILSTM at 0.9496 compared to 0.8827 for the LSTM. Overall, these results indicate that the BILSTM model performs better than the standard LSTM across all evaluation metrics, making it more effective for the classification task.

The results clearly show that the BILSTM model outperforms the standard LSTM model across most evaluation metrics due to its ability to capture context from both directions.

4.4. Confusion Matrix Analysis

The confusion matrix provides a detailed breakdown of classification performance. The results of the proposed model are illustrated in Fig. 2.

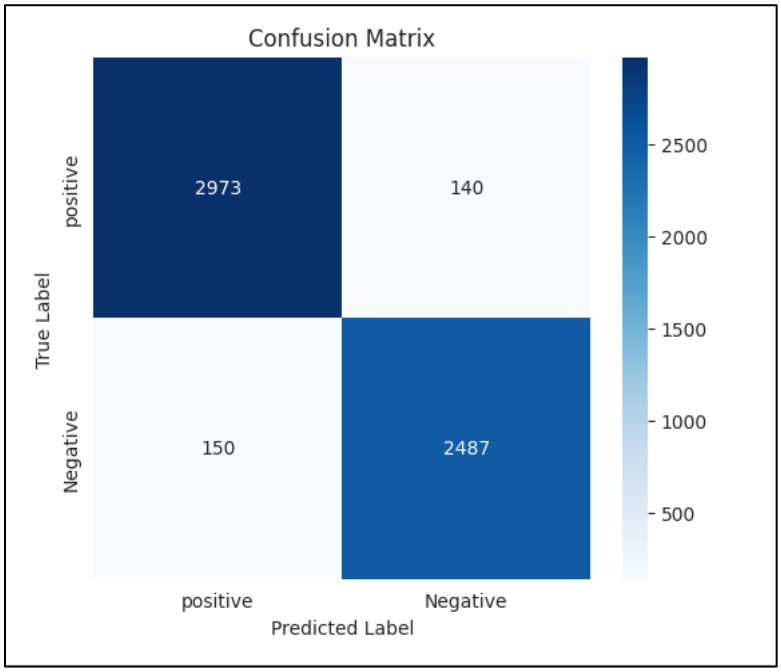

Figure 2 Confusion matrix representing classification results of the proposed BILSTM model

The model demonstrates:

- High true positive and true negative rates, indicating accurate detection of both AI and human text.
- Low misclassification (FP and FN), showing strong reliability of the model.
- Balanced performance across both classes, ensuring consistent prediction without bias.
-

4.5. Training and Validation Analysis

During training, the model showed consistent improvement in accuracy with each epoch, while the loss decreased steadily. The validation accuracy closely followed the training accuracy, indicating that the model generalised well without significant overfitting, as seen in Fig. 3.

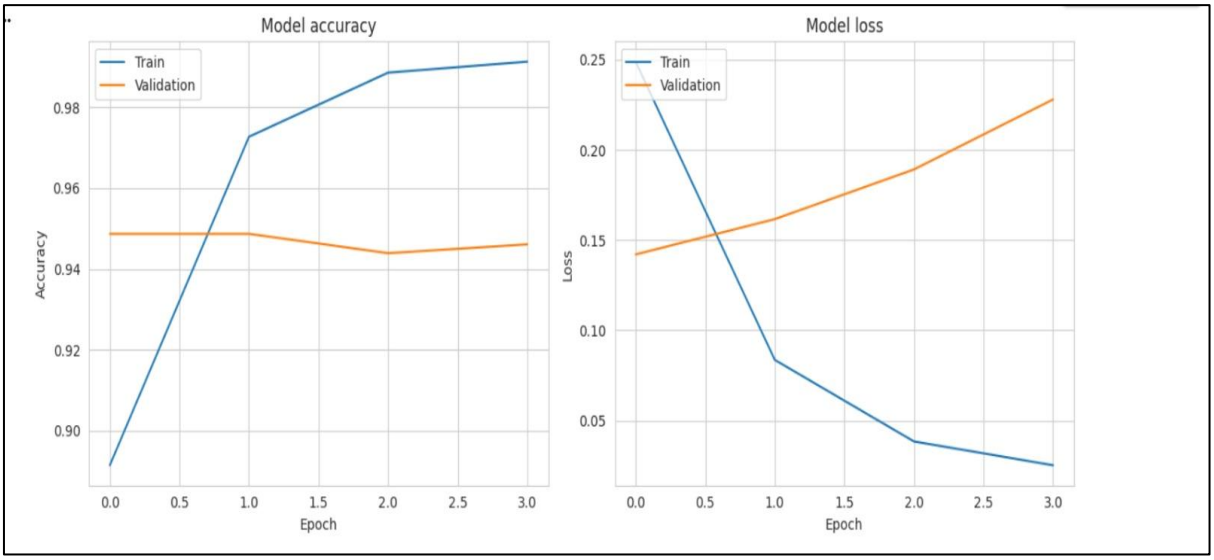

Figure 3 Model Performance Curves Showing Accuracy and Loss During Training

Key Points:

- Training loss decreases gradually as the model learns linguistic signatures.
- Validation accuracy stabilises at **94.96%** after a few epochs.
- No major overfitting was observed despite the depth of the Bilt layers.

5. Discussion

The experimental results clearly demonstrate the effectiveness of the proposed Bilt model for AI-generated text detection. The model successfully captures contextual dependencies and handles complex linguistic structures often found in modern LLM outputs.

Key findings include

- **Improved Context Understanding:** Bilt captures both past and future context, which is essential for identifying the rigid structural patterns of AI text.
- **Robust Performance:** The model achieves a high accuracy of **94.96%**, indicating reliable performance for academic and professional use.
- **Balanced Classification:** Precision and recall values are well-balanced, reducing the risk of false accusations of AI use.
- **Scalability:** The model handled the **28,747** instances efficiently, proving it can be scaled for larger real-world datasets.

Compared to traditional machine learning methods, this deep learning approach significantly improves performance by eliminating the need for manual feature engineering and enabling automatic extraction of complex textual features.

6. Conclusion and Future Work:

This research presented a robust deep learning-based framework for the detection and classification of AI-generated text using Bidirectional Long Short-Term Memory (BILSTM) architectures. The primary objective was to address the growing concerns regarding academic integrity and misinformation by developing a system capable of distinguishing machine-generated content from human writing.

A specialised dataset consisting of **28,747** instances was utilised, providing a balanced distribution of human-authored and AI-generated samples. A comprehensive preprocessing pipeline was applied to clean and standardise the textual data, followed by transformation into numerical representations using tokenisation, padding, and word embedding techniques.

The proposed Bilt model demonstrated superior performance in capturing bidirectional contextual dependencies and structural patterns unique to Large Language Models (LLMs). Experimental results showed that the model achieved a high testing accuracy of **94.96%**, along with optimised precision, recall, and F1-score values. The confusion matrix analysis further confirmed the model's reliability, maintaining a low rate of misclassification across both classes.

The findings of this study highlight the effectiveness of Bidirectional LSTM networks in identifying machine-generated linguistic signatures. By processing text in both directions, the model accurately interprets complex dependencies, making it a valuable tool for content moderation and authenticity verification.

6.1. Future Work

Future research can explore the integration of Transformer-based architectures such as BERT or GPT-based detectors, which have demonstrated state-of-the-art performance in complex natural language understanding tasks. Additionally, extending the system to support multilingual AI detection would enhance its global applicability. Incorporating attention mechanisms and hybrid models, such as combining CNNs with BISTs, may further improve the model's ability to detect localised stylistic anomalies and broader contextual shifts. Finally, the research could shift toward real-time integration for live content moderation platforms.

Compliance with ethical standards

Acknowledgments

The author is thankful to the Department of Computer Science at Government College (Autonomous), Rajahmundry, for their support and encouragement throughout the course of this work**.**

Disclosure of conflict of interests:

The author declares that there are no competing interests

References

[1] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?,” in Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, 2021. doi: 10.1145/3442381.3449820.

[2] S. K. DUVVURI, Applications of Artificial Intelligence Across Domains . Commissionerate of Collegiate Education, Government of Andhra Pradesh , 2026. doi: 10.5281/zenodo.18623057.

[3] OpenAI, “Language Models are Unsupervised Multitask Learners,” OpenAI Blog, 2019, [Online]. Available: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

[4] J. Hirschberg and C. D. Manning, “Natural Language Processing: State of the Art, Current Trends and Challenges,” Comput. Speech Lang., 2019.

[5] D. Bahdanau, K. Cho, and Y. Bengio, “Neural Machine Translation by Jointly Learning to Align and Translate,” arXiv preprint arXiv:1409.2329, 2014.

[6] Pandiri Lavanya, Patinavalasa Durga Prasad, and Suneel Kumar Duvvuri, “Context-Aware Sentiment Classification of Movie Reviews Using Bidirectional LSTM Networks,” Int. J. Sci. Res. Sci. Eng. Technol., vol. 13, no. 2, pp. 159–171, Mar. 2026, doi: 10.32628/IJSRSET261371.

[7] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, and C. Finn, “DetectGPT: Zero-Shot Machine-Generated Text Detection using Probability Curvature,” arXiv preprint arXiv:2301.11305, 2023, [Online]. Available: https://arxiv.org/pdf/2301.11305.pdf

[8] G. Jawahar, M. Abdul-Mageed, and L. V. S. Lakshmanan, “Automatic Detection of Machine Generated Text: A Critical Survey,” arXiv preprint arXiv:2212.09561, 2022, [Online]. Available: https://arxiv.org/pdf/2212.09561.pdf

[9] M. Iyyer, J. Wieting, and K. Gimpel, “Neural Text Generation: A Practical Guide,” arXiv preprint arXiv:2006.05524, 2020, [Online]. Available: https://arxiv.org/pdf/2006.05524.pdf

[10] M. Suja, P. Lavanya, P. D. Prasad, and S. K. Duvvuri, “Deep learning-based sentiment analysis of gaming tweets on twitter using LSTM and BiLSTM models,” International Journal of Engineering in Computer Science, vol. 8, no. 1, pp. 215–222, Jan. 2026, doi: 10.33545/26633582.2026.v8.i1b.269.

[11] A. Graves, “Supervised Sequence Labelling with Recurrent Neural Networks,” University of Toronto, 2013. [Online]. Available: https://www.cs.toronto.edu/~graves/preprint.pdf

[12] I. et al. Sutskever, “Generating Sequences with Recurrent Neural Networks,” arXiv preprint arXiv:1308.0850, 2013.

[13] D. P. Patinavalasa and D. Suneel Kumar, “Scalable Email Spam Detection Using BiLSTM with Large-Scale Hybrid Datasets,” International Journal Of Recent Trends In Multidisciplinary Research, p. 96, Mar. 2026, doi: 10.59256/ijrtmr.20260602016.

[14] Y. LeCun, Y. Bengio, and G. Hinton, “Deep Learning,” arXiv preprint arXiv:1802.06006, 2018.

[15] Y. Goldberg, “Neural Network Methods for Natural Language Processing,” Synthesis Lectures on Human Language Technologies, vol. 10, pp. 1–309, Mar. 2017, doi: 10.2200/S00762ED1V01Y201703HLT037.

[16] Y. Kim, “Convolutional Neural Networks for Sentence Classification,” arXiv preprint arXiv:1408.5882, 2014.

[17] X. Zhang, J. Zhao, and Y. LeCun, “Character-level Convolutional Networks for Text Classification,” arXiv preprint arXiv:1509.01626, 2015.

[18] A. Conneau, H. Schwenk, L. Barrault, and Y. Lecun, “Very Deep Convolutional Networks for Text Classification,” arXiv preprint arXiv:1606.01781, 2016.

[19] S. Hochreiter and J. Schmidhuber, “Long Short-Term Memory,” Neural Comput., vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.

[20] M. Schuster and K. K. Paliwal, “Bidirectional recurrent neural networks,” IEEE Transactions on Signal Processing, vol. 45, no. 11, pp. 2673–2681, 1997, doi: 10.1109/78.650093.

[21] A. Vaswani et al., “Attention Is All You Need,” arXiv preprint arXiv:1706.03762, 2017.

[22] J. Devlin, M. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” NAACL, 2019.

[23] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, and E. Hovy, “Hierarchical Attention Networks for Document Classification,” arXiv preprint arXiv:1606.02393, 2016.

[24] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language Models are Unsupervised Multitask Learners,” OpenAI Technical Report, 2019, [Online]. Available: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

[25] T. B. Brown et al., “Language Models are Few-Shot Learners,” in Advances in Neural Information Processing Systems (NeurIPS), 2020. [Online]. Available: https://arxiv.org/pdf/2005.14165.pdf

[26] R. et al. Zellers, “Grover: A State-of-the-Art Defense against Neural Fake News,” arXiv preprint arXiv:1906.04043, 2019.

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

[28] J. Howard and S. Ruder, “Universal Language Model Fine-tuning for Text Classification,” arXiv preprint arXiv:1801.06146, 2018.

[29] G. Jawahar, B. Sagot, and D. Seddah, “What Does BERT Learn about the Structure of Language?,” arXiv preprint arXiv:1906.04341, 2019.

[30] A. et al. Radford, “Language Models are Unsupervised Multitask Learners,” in ACL Conference, 2020.

[31] T. B. et al. Brown, “Language Models are Few-Shot Learners,” arXiv preprint arXiv:2005.14165, 2020.