International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

Detection of AI-Generated text Using DL Techniques

M. Mahesh[1]*, Mudavath Mallesh[2], Kashimpur Varshitha[2], Jindrala Bhanu Prakash[2], Gajaboina Manaswini[2]

[1]Associate Professor, [2]UGStudent, [1,2]Department of Artificial Intelligence & Machine Learning

[1,2]J. B. Institute of Engineering & Technology (UGC-Autonomous), Moinabad, Hyderabad 500075, Telangana.

*Corresponding author: M. Mahesh (mahesh.m528@gmail.com)

ABSTRACT

This project focuses on the detection of AI-generated text using advanced deep learning (DL) techniques to ensure content authenticity and integrity in the digital era. With the rapid growth of powerful text generation models based on transformer architectures, distinguishing between human-written and machine-generated content has become increasingly challenging. In this work, a robust detection system is developed using transformer-based models trained on a diverse dataset comprising both human-authored and AI-generated text samples. The model leverages contextual understanding, linguistic patterns, and semantic features to accurately classify the text. Various preprocessing techniques and feature extraction methods are applied to improve performance, and the system is evaluated using standard metrics such as accuracy, precision, recall, and F1-score. The proposed approach aims to provide an efficient and scalable solution for identifying AI-generated content, which can be useful in applications such as academic integrity, misinformation control, and content verification.

Key Words: AI-generated text detection, Deep Learning, Transformer models, Natural Language Processing (NLP), Text classification, Machine learning, Feature extraction, Semantic analysis, Dataset, Accuracy, Precision, Recall, F1-score

1. INTRODUCTION

In recent years, the rapid advancement of Artificial Intelligence (AI) has significantly transformed the field of Natural Language Processing (NLP). Modern AI models, especially those based on deep learning and transformer architectures, are capable of generating human-like text with remarkable accuracy and fluency. These models are widely used in applications such as chatbots, content creation, virtual assistants, and automated writing tools. While these technologies offer numerous benefits, they also introduce serious challenges related to authenticity, trust, and misuse of generated content.

One of the major concerns is the difficulty in distinguishing between human-written and AI-

generated text. As AI models become more sophisticated, the generated content often appears natural, coherent, and contextually relevant, making manual detection nearly impossible. This raises critical issues in areas such as academic integrity, fake news detection, plagiarism, and online misinformation. Therefore, there is a growing need for reliable systems that can automatically identify whether a piece of text is generated by a machine or written by a human.

Deep Learning (DL) techniques have emerged as powerful tools for solving complex problems in text analysis and classification. Among these, transformer-based models have gained significant attention due to their ability to understand context and capture long-range dependencies in text. Models such as BERT, GPT, and RoBERTa have shown exceptional performance in various NLP tasks, including text classification, sentiment analysis, and language understanding.

In this project, we focus on developing an efficient system for detecting AI-generated text using deep learning techniques. The proposed approach leverages transformer-based architectures to analyze linguistic patterns, contextual features, and semantic structures present in the text. By training the model on a well-curated dataset consisting of both human-written and AI-generated samples, the system learns to differentiate between the two with high accuracy.

The process begins with data collection and preprocessing, where the dataset is cleaned, tokenized, and prepared for model training. Feature extraction techniques are applied to convert textual data into numerical representations that can be processed by deep learning models. The transformer model is then fine-tuned to classify the input text into appropriate categories.

To evaluate the performance of the model, standard metrics such as accuracy, precision, recall, and F1-score are used. These metrics help in understanding the effectiveness of the model in correctly identifying AI-generated content. The system is also tested on unseen data to ensure its generalization capability and robustness.

The proposed system has practical applications in multiple domains. In the field of education, it can help detect AI-generated assignments and maintain academic honesty. In journalism and social media, it can assist in identifying fake or misleading content. Additionally, it can be integrated into content moderation systems to improve the reliability of online platforms.

In conclusion, the detection of AI-generated text is becoming increasingly important in today's digital world. By utilizing advanced deep learning techniques and transformer models, this project aims to provide an effective and scalable solution to this growing problem. The results of this study contribute to enhancing trust, transparency, and accountability in AI-generated content.

2. LITERATURE SURVEY

The rapid evolution of Artificial Intelligence (AI) and Natural Language Processing (NLP) has led to the development of highly advanced text generation models, which has created a parallel need for reliable detection mechanisms. Several researchers have explored different approaches to identify AI-generated text using machine learning and deep learning techniques.

Early methods for text classification relied on traditional machine learning algorithms such as Naïve Bayes, Support Vector Machines (SVM), and Logistic Regression. These approaches primarily focused on handcrafted features like word frequency, n-grams, and syntactic patterns. Although these models performed reasonably well for basic classification tasks, they lacked the ability to capture deep contextual and semantic relationships in text, making them less effective for detecting sophisticated AI-generated content.

With the introduction of deep learning, models such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks were applied to text classification problems. These models improved the ability to understand sequential data and context within sentences. However, they still faced limitations in handling long-range dependencies and required extensive training time.

A major breakthrough came with the development of transformer-based architectures, which revolutionized NLP tasks. Models like BERT (Bidirectional Encoder Representations from Transformers) introduced bidirectional context understanding, allowing the model to analyze text more effectively. Similarly, GPT (Generative Pre-trained Transformer) models demonstrated powerful text generation capabilities, which also increased the challenge of detection.

Researchers have proposed various transformer-based detection systems that leverage fine-tuned pre-trained models for identifying AI-generated text. These systems analyze linguistic features such as sentence structure, coherence, perplexity, and token distribution to differentiate between human and machine-generated content. Studies have shown that transformer models outperform

traditional and sequential deep learning models in terms of accuracy and generalization.

Another important area of research involves the use of hybrid models that combine deep learning with statistical methods. These models aim to enhance detection performance by integrating multiple features, including lexical, syntactic, and semantic attributes. Ensemble techniques have also been explored to improve robustness and reduce classification errors.

Recent studies have focused on adversarial approaches, where detection models are trained to identify outputs from increasingly sophisticated AI generators. This has led to the development of more resilient detection systems that can adapt to evolving generation techniques. Additionally, researchers are exploring the use of explainable AI (XAI) to make detection decisions more transparent and interpretable.

Datasets play a crucial role in training effective detection models. Various publicly available datasets containing human-written and AI-generated text have been used in research. The quality, diversity, and size of these datasets significantly impact the performance of detection systems.

Despite significant progress, detecting AI-generated text remains a challenging task due to continuous advancements in generation models. Newer models produce highly coherent and contextually rich content, making it harder for detection systems to distinguish between human and machine-generated text accurately.

In conclusion, the literature highlights that while traditional machine learning methods laid the foundation, deep learning and transformer-based models have become the most effective approaches for AI text detection. Ongoing research continues to focus on improving accuracy, robustness, and adaptability of these systems to keep pace with rapidly evolving AI technologies.

3. PROPOSED SYSTEM

The proposed system aims to accurately detect whether a given input text is AI-generated or human-written by using advanced Deep Learning techniques, specifically transformer-based models. The core idea behind this system is to leverage the powerful contextual understanding capability of transformers to analyze and classify text with high precision. Unlike traditional methods that rely on basic statistical features, this system focuses on understanding the deeper semantic and linguistic patterns present in the text.

The system begins with the input stage, where a user provides a piece of text that needs to be analyzed. This text can be from any source, such as an article, assignment, or social media content. Once the input is received, it undergoes a preprocessing phase. In this phase, unnecessary characters, symbols, and noise are removed to clean the text. The cleaned text is then tokenized, which means it is split into smaller units (tokens) that the model can understand. Tokenization is an important step because transformer models work with tokens rather than raw text.

After preprocessing, the tokenized text is passed into a pre-trained transformer model such as BERT or similar architectures. These models are already trained on large datasets and have a strong understanding of language structure and context. In this project, the transformer model is further fine-tuned using a labeled dataset that contains both AI-generated and human-written text samples. This fine-tuning process helps the model learn the subtle differences between the two types of text.

The transformer model processes the input by analyzing relationships between words in the sentence using attention mechanisms. This allows the model to capture context, meaning, and writing style more effectively than older models. It identifies patterns such as repetition, unnatural phrasing, predictability, and coherence, which are often present in AI-generated text.

Once the model processes the input, it passes the learned features to a classification layer. This layer is responsible for making the final decision. It outputs a probability score indicating whether the text is AI-generated or human-written. Based on this score, the system classifies the input into one of the two categories.

To ensure the effectiveness of the system, it is trained and tested using standard evaluation metrics such as accuracy, precision, recall, and F1-score. These metrics help measure how well the model performs in correctly identifying the type of text. The system is also validated on unseen data to ensure it can generalize well to new inputs.

One of the key advantages of this proposed system is its ability to handle complex and context-rich text due to the use of transformer models. It provides better accuracy compared to traditional machine learning methods and can adapt to different types of writing styles. Additionally, the system is scalable and can be integrated into real-world applications such as plagiarism detection tools, academic verification systems, and content moderation platforms.

In conclusion, the proposed system offers a robust and efficient solution for detecting AI-generated text. By combining preprocessing techniques, transformer-based deep learning models, and effective classification methods, the system achieves reliable performance. This approach not only improves detection accuracy but also contributes to maintaining trust and authenticity in digital content.

4. RESULT DESCRIPTION

The developed system for detecting AI-generated text using deep learning techniques, specifically transformer-based models, has produced highly accurate and reliable results. The system successfully classifies input text as either AI-generated or human-written by analyzing contextual and semantic patterns. Based on the output displayed in the web interface, the model predicts that the given input text is **93.23% AI-generated** and **6.77% human-written**, clearly indicating a strong confidence level in identifying machine-generated content. This demonstrates the effectiveness of the transformer model in capturing subtle linguistic differences between human and AI-generated text.

The system interface, as shown in the figure, provides a clean and user-friendly platform where users can input text and obtain instant results. After clicking the “Detect” button, the system processes the input using the trained model and displays the classification result along with probability scores. The inclusion of both numerical values and visual representation (percentage circle) enhances the clarity and usability of the system.

The analysis section of the interface further simplifies the understanding of results by visually representing the prediction using a circular indicator. The system highlights that the text is predominantly AI-generated, making it easier for users to interpret the output without needing technical knowledge. Additionally, separate percentage indicators for AI text and human text provide a detailed breakdown of the prediction.

The training process of the model was carried out using a transformer-based architecture on a GPU-enabled environment (Tesla T4), which significantly improved training efficiency. The training logs show that the model was trained for multiple epochs, and key performance metrics such as training loss, validation loss, accuracy, precision, recall, and F1-score were recorded.

From the training results, it can be observed that the model achieves high performance. For example, in one of the training runs, the model achieved an accuracy of approximately **94.21%**, precision of **90.72%**, recall of **98.57%**, and F1-score of **94.48%**. These values indicate that the model performs well in correctly identifying AI-generated text while maintaining a good balance between precision and recall.

The reduction in training loss across epochs demonstrates that the model is effectively learning patterns from the dataset. Although slight variations in validation loss are observed, the overall performance remains stable, indicating good generalization capability. The use of transformer models allows the system to capture long-range dependencies and contextual relationships within the text, which is crucial for accurate classification.

Another important observation from the results is that the system performs consistently across

International Journal For Advanced Research In Science & Technology
A peer reviewed international journal
www.ijarst.in
ISSN: 2457-0362

different training sessions. Even in another training instance, the model achieved accuracy values above **93%**, confirming its robustness and reliability. The use of GPU acceleration further ensures faster training and efficient handling of large datasets.

The system also demonstrates real-time prediction capability, where users can input text and receive immediate results. This makes the system highly practical for applications such as academic integrity checking, content verification, and detection of AI-generated articles. The fast response time and high accuracy make it suitable for deployment in real-world environments.

The visualization of results through the web interface enhances user experience by providing both textual and graphical outputs. The percentage-based display, along with labeled categories, ensures that even non-technical users can easily understand the results. This improves the accessibility and usability of the system.

Overall, the results confirm that the proposed transformer-based deep learning model is highly effective in detecting AI-generated text. The combination of high accuracy, real-time prediction, and user-friendly interface makes the system a reliable solution for identifying machine-generated content.

Fig 2 Web Interface of AI Text Detection System

The figure shows a modern and interactive web interface where users can enter text and obtain classification results. The interface displays prediction results, confidence scores, and visual indicators, ensuring clarity and ease of use

Fig 3 Training Process using Transformer Model

The figure illustrates the training process of the model using a GPU environment. It includes epoch-wise performance metrics such as training loss, validation loss, accuracy, precision, recall, and F1-score.

Fig 4 Model Performance Metrics

The figure shows the evaluation results of the model, indicating high accuracy and strong classification performance across different metrics.

5. REFERENCES

[1]. Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for

Language Understanding. Proceedings of NAACL-HLT.

[2]. Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training. OpenAI.

[3]. Tom Brown et al. 2020. Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems (NeurIPS).

[4]. Yinhan Liu, Myle Ott, Naman Goyal et al. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint.

[5]. Ashish Vaswani, Noam Shazeer, Niki Parmar et al. 2017. Attention Is All You Need. Advances in Neural Information Processing Systems (NeurIPS).

[6]. Hugging Face. 2023. Transformers Library Documentation. Available: https://huggingface.co/docs

[7]. OpenAI. 2019. GPT-2 Output Dataset and Research Overview. Available: https://openai.com

[8]. Kaggle. 2023. AI vs Human Text Dataset. Available: https://www.kaggle.com

[9]. PyTorch. 2023. PyTorch Documentation. Available: https://pytorch.org

[10]. Scikit-learn. 2023. Machine Learning Library Documentation. Available