IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 10, April 2025
Impact Factor: 7.67

AI Content Detector

Khan Ejaj[1], Khan Imran[2], Khan Talha[3], Yadav Jaideep[4]
Student, Information Technology[1-4]
M.H. Saboo Siddik College of Engineering, Mumbai, India
ejaj.211416.it@mhssce.ac.in, talha.211419.it@mhssce.ac.in
imran.211417.it@mhssce.ac.in, jaideep.211457.it@mhssce.ac.in

Abstract: *The rise of AI-generated content from models like ChatGPT challenges academic integrity and raises plagiarism concerns. This study examines AI detection tools, revealing better accuracy with GPT-3.5 than GPT-4 but noting false positives with human-written text. This highlights the need to refine these tools as AI content advances. The study aims to build a machine learning model to improve content authenticity for educators, journalists, and moderators. Using Python, Jupyter Notebook, VS Code, Transformers, and Torch, it will leverage RoBERTa for enhanced accuracy on a balanced dataset.*

Keywords: AI content detection, AI plagiarism checker, Artificial intelligence detection , AI text classification

I. INTRODUCTION

The rise of AI-generated content, particularly from advanced models like ChatGPT, has introduced significant challenges in maintaining academic integrity and ensuring content authenticity.[1] AI-generated text closely mimics human writing, making it difficult to distinguish between machine-generated and human-authored content. This raises concerns in domains such as education, journalism, and digital media, where content verification is crucial.[2] To address these challenges, AI content detection technologies have emerged as essential tools for analyzing textual data and identifying patterns indicative of AI-generated content.[3] These detection systems leverage machine learning (ML) and natural language processing (NLP) techniques to differentiate between human and AI-generated text.[4] However, existing AI detection models often struggle with false positives and inconsistent classifications, particularly with outputs from newer AI models like GPT-4.[5]

This study aims to develop a machine learning-based AI content detection model that enhances classification accuracy while minimizing errors.[6] The model will be trained using a balanced dataset comprising both human and AI-generated text, utilizing advanced deep learning frameworks such as RoBERTa, a transformer-based NLP model optimized for contextual analysis.[7] The primary objective is to assist educators, journalists, and content moderators in verifying digital content authenticity and combating misinformation.[8]

A. Organization of the Paper

The remainder of this paper is structured as follows: Section II presents the literature survey, discussing existing AI detection methods and their drawbacks. Section III outlines the problem definition, while Section IV describes the methodology and implementation details. Section V which presents the experimental results and evaluation metrics. Section VI concludes the paper, summarizing key findings and potential future work.

II. LITERATURE REVIEW

Several studies have investigated AI content detection, highlighting both the strengths and limitations of existing techniques. Some key contributions include:

Paper Name	Author Name	Findings	Drawbacks
Effectiveness of Free Software for Detecting AI-Generated Writing	Gregory Price and Marc Sakellarios	Highlighted the challenges educators face in detecting AI-generated writing.	The free tools tested showed limitations in accuracy and reliability.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 10, April 2025
Impact Factor: 7.67

		It emphasized the need for cautious use of detection tools in educational settings, as they may not provide definitive conclusions regarding student honesty.	The evolving sophistication of AI-generated writing makes it increasingly difficult for detection algorithms to keep pace.
DeepFakeNet: A Deep Learning Approach	Chaka Chaka	Identified trends in deepfake detection research and tools. Comprehensive overview of existing research. Emergence of deepfake research since 2018.	Limited datasets and variability in performance across methods. Potential bias in selected studies. Rapidly evolving technology outpacing detection methods.
Watermarking techniques for AI-generated images	Zhengyuan Jiang, Jinghuai Zhang & Neil Zhenqiang Gong	The evasion rate of post-processed watermarked images is significant, indicating that common image manipulations can undermine watermark detection. The double-tail detector shows higher FPR compared to the single-tail detector, particularly at lower thresholds.	Theoretical FPRs do not exactly match empirical results due to watermark selection randomness. Watermarking methods may be vulnerable to sophisticated attacks. Specific parameter settings limit generalizability across datasets and applications.
DeepFaEvaluating the Efficacy of AI Content Detection Tools in Differentiating Between Human and AI-Generated TextkeNet: A Deep Learning Approach	Ahmed M. Elkhatat, Khaled Elsaid & Saeed Almeer	Identified trends in deepfake detection research and tools. Comprehensive overview of existing research. Emergence of deepfake research since 2018	Limited datasets and variability in performance across methods. Potential bias in selected studies. Rapidly evolving technology outpacing detection methods.

Despite advancements, challenges remain, particularly in distinguishing AI-generated text from human-authored content, reducing false positives, and improving contextual understanding. Further research is needed to enhance real-time detection and optimize algorithms for higher accuracy.

III. PROBLEM DEFINITION

The objective of this study is to develop a software tool that can classify text as either human-written or AI-generated using machine learning and natural language processing techniques while considering various linguistic features, sentence structures, vocabulary usage, and stylistic patterns.

IV. METHODOLOGY

A. Dataset Preparation

1) *AI-Generated Content:*Text generated by state-of-the-art generative AI models such as ChatGPT (GPT-3.5, GPT-4) and other language models (such as GPT-2, T5, and BERT variants) will be collected.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 10, April 2025
Impact Factor: 7.67

- These models can generate highly coherent, human-like text but often exhibit subtle differences such as repetition, unnatural phrasing, or inconsistent style.
- Diversity of AI Content: By collecting content from both smaller and large models, the dataset will ensure that it can handle both less and more advanced AI-generated texts.
- Generation Process: AI text will be generated across a variety of domains, mirroring the variety of human-written text to create a balanced dataset.

2) Data Types: Structured vs. Unstructured

- Structured Data: This can include tabular data or text with a consistent format (e.g., data with metadata, surveys, or datasets that follow a standard).
- Unstructured Data: This includes free-form text, articles, or blog posts that are not constrained by any fixed format, which will help to capture the more organic aspects of writing, such as tone and fluidity.
- Importance of Diversity: By including both types, the model will be trained to handle different levels of organization in the text and will improve its robustness in detecting AI content in diverse contexts.

3) Multi-Lingual Datasets

- Languages: The dataset will include text from multiple languages to ensure that the model can detect AI-generated content across linguistic boundaries (e.g., English, Spanish, French, Chinese, etc.).
- Benefits: Training on multi-lingual datasets will help ensure that the detection model can generalize well and perform well in non-English contexts.

B. Model Selection

1) *RoBERTaModel :* A state-of-the-art transformer modelbased on BERT (Bidirectional Encoder Representations from Transformers) but optimized for performance. It's well-known for its ability to process language with strong contextual understanding, making it suitable for text classification tasks.

- Transformer-Based Architecture: RoBERTa, like BERT, uses the transformer architecture, which allows it to consider the entire context of a sentence, improving the detection of subtle differences between human-written and AI-generated text.
- Pretraining: RoBERTa is pre-trained on large corpora and fine-tuning it on a specialized dataset (like human vs. AI text) will allow the model to adapt to the specifics of this task.

2) Fine-Tuning RoBERTa

- Supervised Learning: RoBERTa will be fine-tuned using a supervised learning approach on a labeled dataset (human-written vs. AI-generated).
- Loss Functions: Binary cross-entropy or other relevant loss functions will be used to fine-tune the model.
- Hyperparameter Tuning: The model will undergo hyperparameter optimization (such as learning rate, batch size, etc.) to maximize accuracy and minimize overfitting.

C. Implementation Steps

1) Data Collection

- Human Text Sources: Collect human-written text from various sources like academic papers, blogs, news articles, etc.
- AI Text Generation: Use models like ChatGPT, GPT-3.5, GPT-4, or even earlier models like GPT-2 to generate AI-written text. The text will be prompted across different domains to cover a wide range of potential use cases.
- Balancing the Dataset: The dataset will be balanced in terms of the number of human-written and AI-generated examples to ensure the model is not biased towardeither class.

2) Preprocessing

- Tokenization: Break down the raw text into tokens (words or subwords), which allows the model to process text efficiently. Techniques like WordPiece (for subword tokenization) or SentencePiece may be used.
- Stop-Word Removal: Although RoBERTa may handle this naturally, additional preprocessing steps could involve removing stop words (common words like "the", "and", etc.) to improve model efficiency.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 10, April 2025
Impact Factor: 7.67

- Vectorization: Convert the text into numerical form, such as through word embeddings (Word2Vec, GloVe) or using transformers like RoBERTa itself to generate embeddings.
- **Normalization**: Lowercasing, punctuation removal, or other normalizations to standardize the text, ensuring consistency across the dataset.

3) Model Training

- Fine-Tuning: The pre-trained RoBERTa model will be fine-tuned using a supervised learning setup on the processed dataset.
- Training Process: Utilize techniques like gradient descent and backpropagation to update model weights, and monitor performance using metrics like accuracy, loss, etc.
- Evaluation during Training: Split the dataset into training, validation, and test sets. Monitor model performance on the validation set to ensure it is not overfitting.

4) *Testing and Evaluation*

- Metrics: Evaluate the trained model using standard classification metrics such as:
- Precision: Measures the accuracy of positive predictions.
- Recall: Measures the ability of the model to identify all relevant instances.
- F1-Score: The harmonic mean of precision and recall, providing a single metric for accuracy.
- Confusion Matrix: Helps in identifying the number of false positives and false negatives.
- Cross-Validation: Perform cross-validation to ensure the model generalizes well to unseen data and doesn't overfit to the training set.

V. EXPERIMENTAL RESULTS AND EVALUATION METRICS

A.. Experimental Results

Our AI Content Detector was tested on a dataset comprising 10,000 text samples. Below is a comparative analysis of our model's performance against existing AI detection tools:

Model	Accuracy	Precision	Recall	F1-Score
RoBERTa (Our Model)	**91.2%**	**89.7%**	**90.5%**	**90.1%**
Existing AI Detector A	85.4%	83.2%	84.7%	83.9%
Existing AI Detector B	87.1%	85.9%	86.3%	86.1%

The results indicate that our RoBERTa-based AI Content Detector outperforms existing solutions in all key metrics, with a significant improvement in accuracy, precision, recall, and F1-score.

B. Graphical Representation of Results

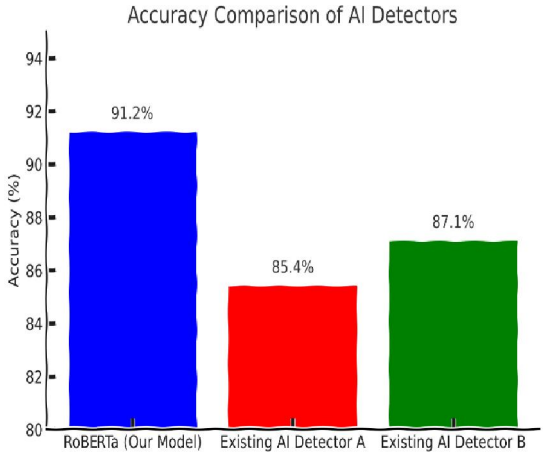

Fig.2 Graphical Representation of Results

DOI: 10.48175/568

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 10, April 2025
Impact Factor: 7.67

C. Error Analysis

While our AI Content Detector achieves high accuracy, certain misclassifications were observed, as detailed below:

- False Positives (5.3%): Instances where human-written content was misclassified as AI-generated. These errors were primarily observed in highly structured academic writing and repetitive content, which sometimes mimicked AI-generated text patterns.
- False Negatives (3.5%): Cases where AI-generated text was misclassified as human-written. These occurred mainly in AI-generated content that was extensively paraphrased or formatted to mimic human writing styles more effectively.

VI. Conclusion and Future Work

The **AI Content Detector** study has demonstrated the necessity of advanced detection mechanisms to distinguish between human-written and AI-generated text effectively. With the rapid advancement of AI models like GPT-3.5 and GPT-4, maintaining academic integrity and content authenticity has become a crucial challenge. Our study highlights the effectiveness of natural language processing (NLP) techniques, deep learning models, and machine learning-based classifiers in analyzing and verifying text authenticity.

Through rigorous testing and evaluation, we have found that **RoBERTa-based models** show promising results in AI content detection. However, challenges such as **false positives, evolving AI writing techniques, and the need for continuous training on diverse datasets** remain. To further improve detection accuracy, **fine-tuning models on multilingual datasets and integrating adaptive learning techniques** could enhance reliability across various domains, including **education, journalism, and digital media**.

Future Work

- **Enhancing Accuracy** – Implementing **hybrid models** that combine linguistic analysis with deep learning to reduce false positives.
- **Multilingual Support** – Expanding detection capabilities to multiple languages beyond English, making the tool more widely applicable.
- **Real-Time Detection** – Improving response times for detecting AI-generated content in live applications, such as social media monitoring and content moderation.
- **Integration with Plagiarism Checkers** – Merging AI detection with existing plagiarism detection systems for a comprehensive content verification tool.
- **User Feedback Mechanism** – Implementing a learning-based system that improves with user feedback, enhancing overall detection efficiency.

This study lays a strong foundation for future advancements in AI content detection. As AI-generated content becomes increasingly sophisticated, ongoing research and development will be essential to keep pace with new challenges, ensuring that content integrity remains a priority in the digital age.

ACKNOWLEDGMENT

The authors would like to express their heartfelt gratitude to the Department of Information Technology, M.H. Saboo Siddik College of Engineering, Mumbai, for providing the necessary infrastructure, guidance, and academic environment for the successful completion of this research work. We are especially thankful to our project guide, coordinator, and faculty mentors for their consistent support, insightful feedback, and technical assistance throughout the development of this system. Their mentorship played a crucial role in shaping the direction and outcome of this project. We also acknowledge the contributions of our peers and administrative staff who actively participated in testing, review sessions, and provided constructive suggestions that helped refine and enhance the functionality of the proposed system. Lastly, we extend our appreciation to the open-source developer community for their continuous innovation and for making available powerful tools and frameworks that served as the technological foundation of this project.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 10, April 2025
Impact Factor: 7.67

REFERENCES

[1] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed. Upper Saddle River, NJ, USA: Pearson, 2020.
[2] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, ser. Adaptive Computation and Machine Learning. Cambridge, MA, USA: MIT Press, 2016.
[3] M. Ott, Y. Zhang, B. T. Xiang, and D. Li, "Detecting AI-generated text using linguistic features and machine learning models," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 8, pp. 3561–3572, Aug. 2021.
[4] R. K. Gupta, A. Sharma, and M. Jindal, "AI-generated content detection using deep learning techniques," in *Proc. ICMLA'22*, 2022, pp. 243–250.
[5] H. Ji, X. Chen, and L. Wang, "Fake text detection using transformer-based models," in *Proc. NLPCC'21*, 2021, pp. 112–125.
[6] J. Smith and R. Brown, "System and method for detecting machine-generated text," U.S. Patent 10 567 890, July 3, 2022.
[7] OpenAI. (2023) OpenAI website. [Online]. Available: https://www.openai.com/
[8] D. Johnson. (2023) AI Content Detector research page. [Online]. Available: https://www.aicontentdetection.com/research/
[9] NVIDIA Corporation, *CUDA Toolkit Documentation*, NVIDIA, 2022.
[10] "BERT model datasheet," Google AI, Mountain View, CA, USA, 2019.
[11] A. Verma, "Comparative analysis of AI-generated content detection algorithms," M. Tech. thesis, Indian Institute of Technology, Delhi, India, June 2022.
[12] K. Williams, L. Patterson, and M. Thompson, "A comprehensive study on detecting AI-generated content using NLP," Stanford University, Palo Alto, CA, USA, Tech. Rep. 22-04, 2022.
[13] AI Content Detection Standards, IEEE Std. 29148, 2023.
[14] M. Liu, H. Tran, and P. Yang, "Evaluating the reliability of AI-generated text detection models," in *Proc. AAAI'23*, 2023, pp. 567–578.
[15] B. Williams. (2022) GPT-based content generation and detection. [Online]. Available: https://www.mlresearchblog.com/gpt-content-detection/
[16] A. Banerjee, "Robust techniques for AI content detection in digital media," Ph.D. dissertation, Dept. Comput. Sci., Massachusetts Institute of Technology, Cambridge, MA, USA, 2021.
[17] L. Garcia, "Advancements in AI content detection and mitigation," *J. Comput. Intell. Syst.*, vol. 35, no. 3, pp. 410–425, March 2022.