AI Content Detector

Khan Talha
Khan Imran
Jaideep Yadav
Khan Ejaj

In the Guidance of Er. ShridnidhiGIndi

CONTENTS

01

Introduction to AI Content Detector

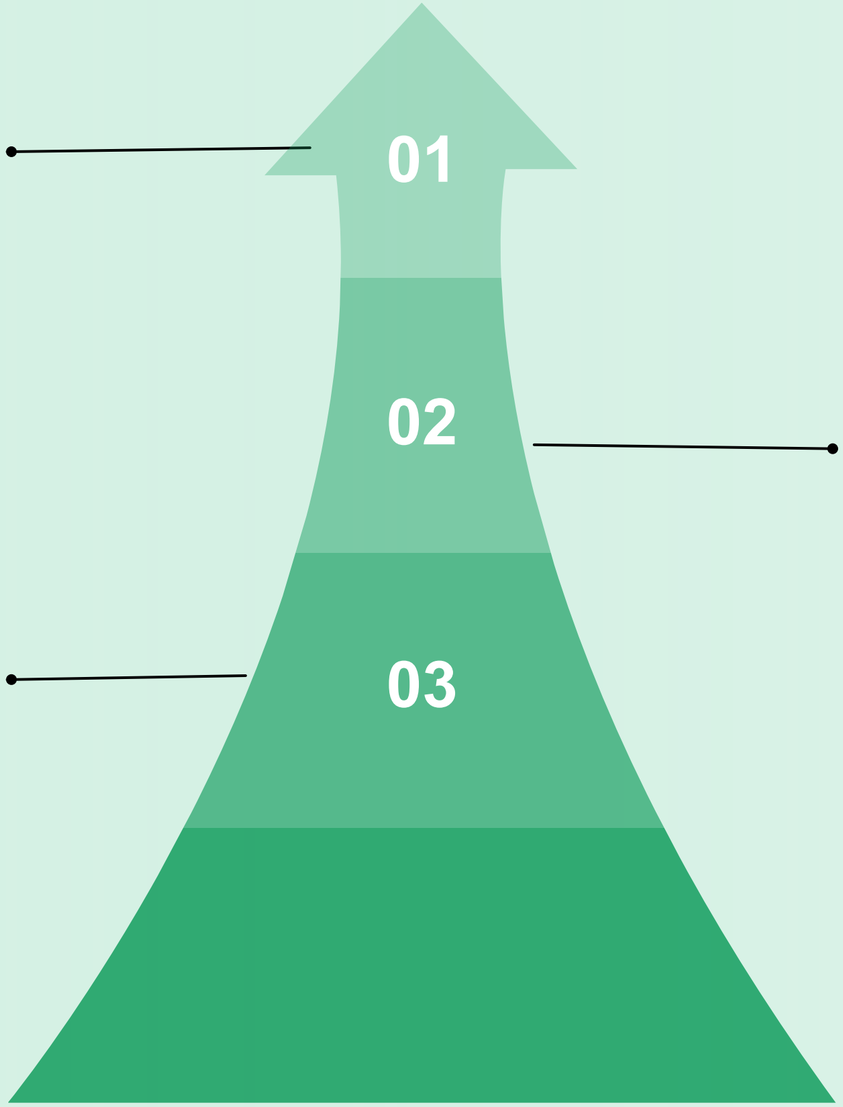
Purpose and Scope

Objective

The objective is to create a machine learning model that accurately differentiates between human and AI-generated content, improving content reliability and authenticity.

Importance in various fields

This model is crucial in fields like education, where integrity and originality are paramount, and journalism, where verifying sources is essential.



Target users

The primary users include educators, journalists, and content moderators who require tools to assess content authenticity and mitigate misinformation.

Applications

Education sector

In education, the model can be utilized to detect plagiarism and ensure that students submit original work, maintaining academic integrity.

Journalism practices

Journalists can use the model to authenticate sources and enhance credibility by distinguishing between human reporting and AI-generated stories.

Content moderation

Content moderators can employ this tool to filter out misleading or automatically generated posts, enhancing the quality of online discourse.

LITERATURE SURVEY

PAPER	YEAR	PURPOSE	TECHNIQUE	RESULT	PARAMETERS	FINDINGS	DRAWBACKS
Effectiveness of Free Software for Detecting AI-Generated Writing	2023	To explore the effectiveness of freely available AI detection software in identifying AI-generated content in student assignments and to assess its implications for teacher evaluation.	• Manual analysis of student writing assignments. • Comparison of results from five different free AI writing detectors.	• The findings revealed limitations in the effectiveness of free AI detection tools, indicating that they may not reliably identify AI-generated content.	• Participants: Japanese university students writing in English as a Foreign Language (EFL). • Tools: Five different free AI writing detection software.	• Highlighted the challenges educators face in detecting AI-generated writing. • It emphasized the need for cautious use of detection tools in educational settings, as they may not provide definitive conclusions regarding student honesty.	• The free tools tested showed limitations in accuracy and reliability. • The evolving sophistication of AI-generated writing makes it increasingly difficult for detection algorithms to keep pace.
DeepFakeNet: A Deep Learning Approach	2019	• To survey existing literature on deepfake detection. • Validate selected studies on deepfake detection. • Analyze empirical evidence in deepfake detection.	• Various detection techniques (e.g., machine learning, deep learning). • Quality assessment criteria. • Data extraction and synthesis.	• Performance evaluation of detection methods. • Finalized 91 research articles and 21 reviews. • Accumulated 112 studies.	• Measurement metrics (e.g., accuracy, precision, recall). • Criteria for inclusion. • Publication period, methods, datasets	• 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.

LITERATURE SURVEY

PAPER	YEAR	PURPOSE	TECHNIQUE	RESULT	PARAMETERS	FINDINGS	DRAWBACKS
watermarking techniques for AI-generated images	2023	The study aims to investigate the robustness of watermarking methods against common post-processing techniques that can be used to evade detection of AI-generated images.	• Watermarking methods: HiDDeN and UDH. • Post-processing methods: Gaussian blur, Brightness/Contrast adjustments, JPEG compression, and Gaussian noise. • The study also introduces an adversarial post-processing method called WEvade-W-II.	• The study finds that existing post-processing methods can effectively evade watermark detection, and the proposed WEvade-W-II outperforms these traditional methods in terms of evasion rate while maintaining image quality.	• For Gaussian blur: Kernel size s=5 and varying standard deviation σ. • For Brightness/Contrast: Parameters a (scaling factor) and b=0.2 (offset). • Watermark lengths: 30-bit for HiDDeN and 256-bit for UDH. • Detection thresholds τ for evaluating the performance of detectors.	• 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.
Evaluating the Efficacy of AI Content Detection Tools in Differentiating Between Human and AI-Generated TextkeNet: A Deep Learning Approach	2023	• To assess the ability of AI detection tools to accurately identify AI-generated text from human-written content. • To explore the challenges AI-generated content poses to academic integrity.	• Various detection techniques (e.g., machine learning, deep leaUsed 15 paragraphs generated by ChatGPT Models 3.5 and 4 on cooling towers in engineering. • Included five human-written paragraphs as control samples.	• AI detection tools identified GPT 3.5-generated content more accurately than GPT 4 content. • Tools showed inconsistencies with human-written control responses, producing false positives.	• Measurement metrics (e.g., accuracy, precision, recall). • Criteria for inclusion. • Publication period, methods, datasets	• Identified trends in deepfake detection research and tools. • Comprehensive overview of existing research. • Emergence of deepfake research since 2018.	• Potential bias in selected studies. • Rapidly evolving technology outpacing detection methods.

02

Tools and Technologies

Tools and Technology

Software Used

- **Python:** simple syntax, powerful libraries for efficient machine learning development.
- **Jupyter Notebook:** easy prototyping, interactive data visualization, and seamless sharing.
- **VS Code:** robust, extensible environment for efficient development, debugging, and collaboration.

Libraries:

- **Transformers:** A library by Hugging Face for state-of-the-art natural language processing.
- **Torch:** The core library for PyTorch, used for machine learning and deep learning models.
- **Gunicorn:** A Python WSGI HTTP server for deploying web applications, useful for serving your models in production environments.

Models:

- **RoBERTa:** Robustly Optimized BERT Pretraining Approach is a variant of BERT that improves performance by training longer, with larger batches, and removing the Next Sentence Prediction task.

03

Model Overview

Model Choice

- **Introduction to RoBERTa**

RoBERTa, a variant of BERT, is designed for enhanced performance in tasks involving natural language understanding, enabling nuanced comprehension of context.

- **Why RoBERTa?**

Its ability to analyze text at a deeper level makes RoBERTa especially suitable for distinguishing between the subtleties of human versus AI-generated content.
For Example:
"The bank is situated on the _____ of the river.

Why Use Roberta?

- **Linguistic Features**

AI Content: Climate change is a critical issue. It affects weather patterns. Solutions include renewable energy. Governments must act nowSimplicity: Shorter sentences, basic vocabulary (e.g., "talk to people far away," "play games"), and conversational tone.

Subjectivity: Personal advice ("listen to our parents") and emotional cues ("enjoy nature").

Subtle Errors: Less precise phrasing (e.g., "waste time" vs. "raise concerns about productivity loss").

- **Structural Patterns**

Human Essay:

Narrative Flow: Casual structure with abrupt topic shifts (e.g., jumping from emergencies to homework).

Repetition: Reiterates basic ideas ("use mobile phones wisely") without elaboration.

AI Essay:

Logical Progression: Each paragraph introduces a distinct theme (functionality → apps → sensors → societal debates).

Balance: Neutral, comprehensive coverage of pros/cons (e.g., "versatility" vs. "privacy concerns").

04

Data Collection

Types of Data

01

Human-generated content

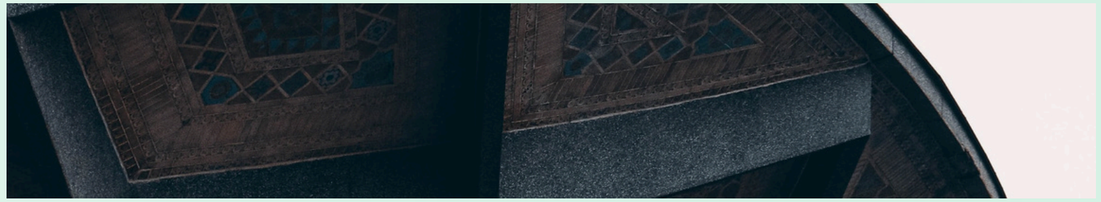
Human-generated content is sourced from blogs, articles, and academic papers, representing authentic expressions and thought processes of individuals.

02

AI-generated content

AI-generated content is collected from automated writing tools and other generative models, allowing for a balanced dataset to train and validate the model.

Challenges and Solutions



Common challenges

Common challenges include dealing with imbalanced datasets and overfitting, which may hinder the generalizability of the model to unseen data.

Mitigation strategies

Mitigation strategies involve employing techniques like data augmentation and regularization methods to enhance robustness and prevent model overfitting.

06

Future Directions

Long-term Goals

Vision for model evolution

The long-term vision includes evolving the model to adapt to new trends in AI content generation and maintain relevance in a rapidly changing digital landscape.

Addressing ethical considerations

Addressing ethical considerations involves ensuring that the model is designed and deployed with fairness and transparency, minimizing biases in content detection.

Conclusion:

The AI Content Detector represents a crucial advancement in verifying the authenticity of information. By effectively distinguishing between human and AI-generated content, we can uphold academic integrity, enhance journalistic credibility, and improve online discourse.

As we navigate an increasingly digital world, tools like this are essential for combating misinformation and fostering trust. Thank you for your attention, and we look forward to shaping the future of content verification together!

Thank you