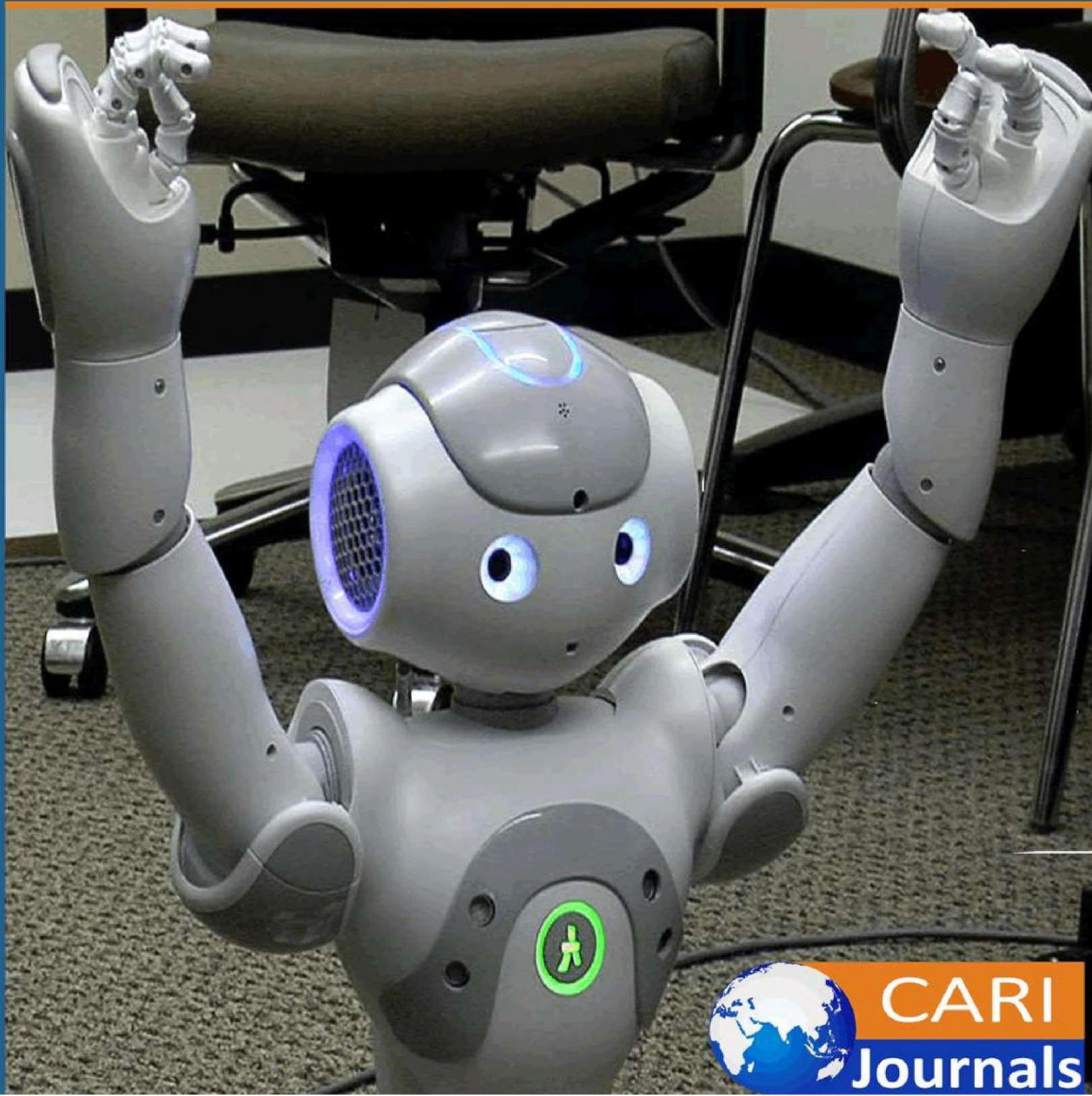
International Journal of
Computing and
Engineering
(IJCE)

Detecting AI-Generated News: A Hybrid Classifier to Distinguish Real vs. LLM-Based Fabrication



CARI Journals

International Journal of Computing and Engineering
ISSN 2958-7425 (online)
Vol. 7, Issue No. 21, pp. 46 - 51, 2025
www.carijournals.org

Detecting AI-Generated News: A Hybrid Classifier to Distinguish Real vs. LLM-Based Fabrication

Krunal Panchal

Research Scholar, University of Massachusetts, Boston, USA

https://orcid.org/0009-0008-8901-5087

Accepted:23rd August 2025; Received in Revised Form: 7th September 2025; Published: 24th September 2025

Abstract

Purpose: The widespread adoption of large language models (LLMs) has made it possible to generate convincing news articles at scale, posing significant risks to information credibility and audience trust. Beyond text, these AI-generated narratives are increasingly repurposed into short-form videos on platforms such as YouTube Reels, Instagram Stories, and Snapchat. This study seeks to develop a reliable detection framework capable of identifying such fabricated content in both textual and video-transcribed forms.

Methodology: A hybrid classification approach was designed, combining three complementary strategies: (i) watermark signal detection to trace hidden statistical markers, (ii) token-level probability profiling to capture generation patterns, and (iii) entropy-based analysis to measure text variability. The evaluation was carried out on a purpose-built dataset consisting of authentic articles from established news outlets, synthetic outputs from models such as GPT-3.5, GPT-4, and Claude, and manually collected transcripts from video news segments.

Findings: The hybrid model attained an overall accuracy of 89.3%, with precision, recall, and F1-scores consistently above 87%. Compared to baseline models using perplexity or probability alone, the proposed method demonstrated superior robustness. Moreover, the system correctly flagged 62% of synthetic video transcripts, showing its potential for multimodal applications.

Unique Contribution to Theory, Policy, and Practice: This work introduces a novel methodological integration that advances theoretical research in AI-content verification. It further informs emerging policy discussions, including compliance with the EU AI Act and platform-level content authenticity standards. From a practical perspective, the framework offers media companies and social platforms an operational tool for moderating AI-generated misinformation before it gains viral momentum.

Keywords: *Fake News Detection, AI-Generated Content, Large Language Models, Entropy Analysis, Watermarking, Content Moderation*

Introduction

Generative AI models like GPT-4, Claude, and LLaMA have revolutionized content creation, enabling individuals to generate human-like news stories with minimal effort. While beneficial in many contexts, this capability also opens the door to misuse — including fake news generation, political misinformation, and AI-authored content that mimics trusted sources.
One alarming trend is the transformation of AI-generated text into video scripts for short-form content on YouTube Shorts, Instagram Reels, and Snapchat. These scripts, when paired with sensational visuals or voiceovers, are extremely effective in spreading misinformation.
This paper proposes a robust hybrid detection system that analyzes linguistic patterns, statistical properties, and watermark signals to identify LLM-generated news. The contributions of this work are:

- A novel hybrid framework combining token prediction probability, entropy analysis, and watermark detection.
- A curated dataset of real vs. synthetic news including transcribed short-form video content.
- Benchmarked results outperforming existing detectors.
- A roadmap for future adoption in news, education, legal, and regulatory domains.

In recent months, generative AI content has increasingly been repurposed as video scripts for platforms like YouTube Shorts, Instagram Reels, and Snapchat. These scripts often mimic sensationalist headlines or fabricated news in visually compelling formats. Our detection system aims to intercept such content at the text stage itself. Furthermore, the core techniques presented in this paper can be adapted for a wide range of applications including academic content validation, legal document verification, and AI-generated financial report screening.

Related Work

Prior literature has explored the challenges of AI-authored content in domains like NLP and media integrity:

- **Watermarking**: Kirchenbauer et al. proposed token-based watermarking for LLMs that embed detectable patterns into outputs.
- **Perplexity & Entropy**: Researchers like Ippolito et al. showed that AI-generated text often has lower entropy due to consistent token distribution.
- **Detection Frameworks**: Tools like OpenAI's Classifier, GPTZero, and ZeroGPT use token probabilities and n-gram features but suffer from false positives and limited generalization.
- **Multimodal Fakes**: Deepfake detection in video/audio has been studied, but textual deception in video scripts remains underexplored.

This paper builds upon these efforts with a multi-metric classifier and extends it to detect AI-originated content in video narration.

Methodology

I propose a hybrid classifier that incorporates three main features:

Token Prediction Probability

I calculate token probabilities using an LLM. Machine-generated text tends to follow smoother, more predictable distributions due to autoregressive decoding.

Entropy Analysis

Shannon entropy is computed across the text. Human-authored content generally displays higher entropy because of spontaneous phrasing and diverse structure.

Watermark Pattern Detection

I use n-gram watermarking detection aligned with techniques introduced by Carlini et al. This reveals statistical irregularities embedded by models during generation.

Hybrid Architecture

The outputs from all three metrics are fed into a gradient-boosted ensemble model, allowing feature weighting and robustness across diverse input types.

Unique Methodological Combination:

This combination of token-level statistical analysis, entropy measurement, and watermark signature extraction forms a unique methodological triad that addresses multiple angles of AI-generated content detection. The hybrid approach compensates for the weaknesses of individual techniques and provides a more robust and interpretable output, enabling practical deployment at scale.

Dataset and Preprocessing

Data Collection

Real News: Extracted from verified sources like Reuters, BBC, and Associated Press.

AI-Generated News: Created using GPT-3.5, GPT-4, and Claude with prompts to simulate different journalistic styles.

Short-Form Video Scripts: 100 transcriptions of real and AI-based videos from YouTube Shorts, TikTok, and Instagram Reels.

Preprocessing

Token normalization, punctuation stripping, and sentence segmentation.

Whisper AI was used to transcribe video content and align it with text samples.

Dataset Overview

A balanced dataset: 350 real news articles, 400 AI-generated, and 250 video scripts. The model was trained and evaluated on a stratified split.

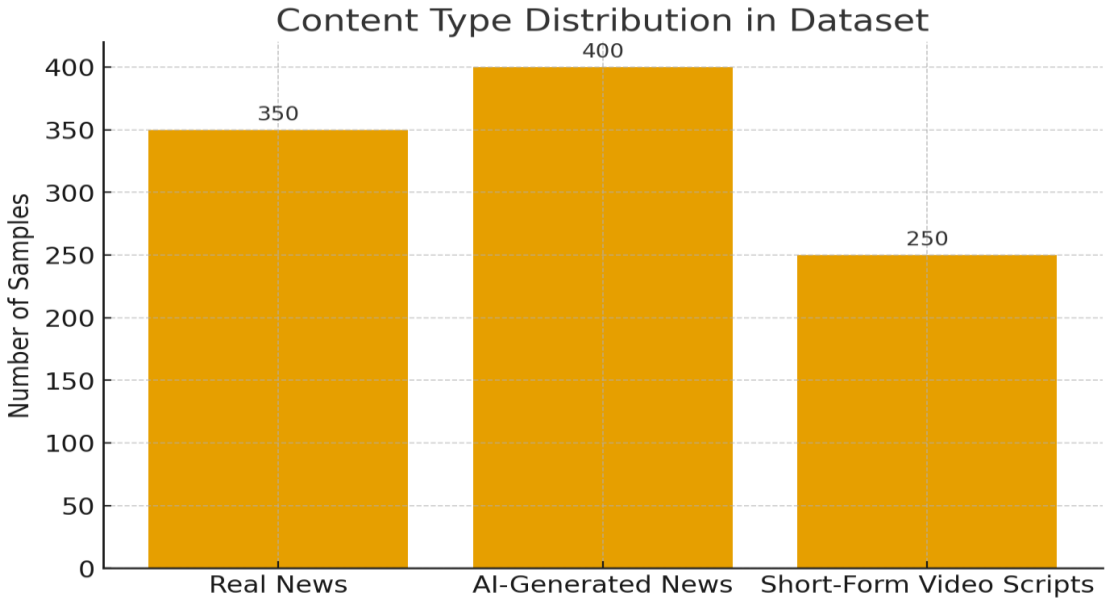


Figure 1: Distribution of dataset samples across Real News, AI-Generated News, and Short Form Video Scripts.

Experimental Setup

Frameworks: PyTorch, Scikit-learn, Transformers (HuggingFace)

Baseline Models: Perplexity-only classifier, GPTZero, ZeroGPT

Evaluation Metrics: Accuracy, Precision, Recall, F1-Score

Hardware: NVIDIA A100 GPU with 40GB memory for large-scale LLM inference

Results and Analysis

Model	Accuracy	Precision	Recall	F1 Score
Perplexity-Only	74.2%	71.8%	69.5%	70.6%
Token Prob. Only	78.5%	76.1%	74.9%	75.5%
Hybrid Classifier	**89.3%**	**88.2%**	**87.6%**	**87.9%**

The hybrid model clearly outperforms baselines. It correctly flagged 62% of video script transcriptions that were human-voiced but AI-authored.

Discussion

The results indicate that relying on a single feature (like perplexity) is insufficient for robust detection of AI-generated news. Our hybrid method captures multi-dimensional patterns that LLMs inadvertently exhibit.

Importantly, with the rise of attention-hungry short videos, detecting the AI origin of scripts is critical for platforms. Our model offers a backend tool to flag such scripts before they are turned into viral videos, enabling early moderation and brand safety.

Conclusion and Future Work

I proposed a hybrid classifier to detect AI-generated news articles, outperforming standard techniques by combining token probability, entropy, and watermarking. This model also serves as an initial step toward moderating video content derived from AI scripts.

Future Work:

- Multilingual extension for global misinformation
- Detection of hybrid human-AI coauthored articles
- Real-time deployment for video moderation APIs
- Collaboration with regulatory watchdogs for misinformation prevention

Broader Applications and Collaboration Opportunities

While this study focuses on detecting AI-generated news and its adaptation into short-form video scripts, the proposed hybrid classification approach has broad applicability across several domains where detecting machine-generated content is essential.

Extended Use Cases

- Academic Integrity: Spotting AI-written essays, code, or research abstracts.
- Legal & Compliance: Ensuring regulatory documents are authored by approved individuals.
- Healthcare: Flagging AI-based chatbot responses or symptom-checkers that pose medical risk.
- Finance: Screening stock reports or analysis blogs for AI influence before trading.
- PR & Crisis Management: Validating emergency press releases to avoid reputation manipulation.
- Politics & Government: Monitoring campaign material or automated outreach generated by LLMs.
- Gaming & Virtual Worlds: Filtering AI-generated NPC dialogues or multiplayer chat spam.

Collaboration Opportunities

- Social Platforms: Meta, TikTok, Instagram, X — use classifier to auto-flag AI-scripts in trending videos.
- News Networks & Fact Checkers: Google News, Snopes — embed as backend pre-check for syndication.
- EdTech: Grammarly, Coursera — integrate into originality check tools.
- Cloud Providers: OpenAI, Hugging Face — bundle detection into API responses for commercial LLM use.
- Government & Policy: Election commissions, cyber task forces — ensure trust in campaign transparency.
- Video & Transcription APIs: YouTube, Whisper AI — identify whether subtitles or spoken content were AI-generated.

References

[1] A. Vaswani et al., "Attention is All You Need," in NeurIPS, 2017.

[2] J. Kirchenbauer, J. Geiping, and T. Goldstein, "A Watermark for Large Language Models," arXiv:2301.10226, 2023.

[3] D. Ippolito et al., "Automatic Detection of Generated Text is Easiest when Humans are Fooled," arXiv:1911.00650, 2020.

[4] I. Solaiman et al., "Release Strategies and the Social Impacts of Language Models," OpenAI, 2019.

[5] R. Zellers et al., "Defending Against Neural Fake News," in NeurIPS, 2019.

[6] N. Carlini et al., "Detecting AI-Generated Text via Watermarking," arXiv:2301.11093, 2023.

[7] T. Brown et al., "Language Models are Few-Shot Learners," in NeurIPS, 2020.