

Detection of AI-Generated Misinformation and Deepfake Content Using Artificial Intelligence

Shubham Jaidas Bhise¹, Aarti Sanjay Gawai²

¹ Student, Department of Information Technology and Computer Science,

D.G Ruparel College of Arts, Science and Commerce, Mumbai, Maharashtra, India
shubhambhise4063@gmail.com

² Assistant Professor, Department of Information Technology and Computer Science,
D.G. Ruparel College of Arts, Science and Commerce, Mumbai, Maharashtra, India

Abstract - The rapid advancement of Artificial Intelligence (AI) and generative AI technologies has increased the creation of realistic text, images, audio, and video, making AI-generated misinformation and deepfake content a growing concern on social media. Traditional detection methods often focus on a single type of media and face challenges when dealing with multimodal content, evolving generation techniques, and real-world social media conditions. This study investigates AI-based approaches for detecting manipulated and AI-generated content using Machine Learning (ML), Deep Learning (DL), Natural Language Processing (NLP), computer vision, and multimodal detection techniques. The proposed methodology includes dataset selection, data preprocessing, model implementation, and performance evaluation using standard metrics such as accuracy, precision, recall, F1-score, and Receiver Operating Characteristic–Area Under the Curve (ROC-AUC). The study also examines existing research gaps, including cross-dataset generalization, feature fusion, explainability, and robustness against emerging AI-generation methods. By evaluating suitable detection approaches across different forms of digital content, the research aims to contribute to more reliable automated content verification systems that can improve information integrity and support safer social media environments.

Key Words: Artificial Intelligence (AI), AI-Generated Misinformation, Deepfake Detection, Machine Learning (ML), Natural Language Processing (NLP), Multimodal Detection

1. INTRODUCTION

Artificial Intelligence (AI) is changing the way digital information is generated, shared, and consumed. Social media platforms have made it possible to publish text, images, audio,

and video within seconds, allowing information to reach large audiences with very little delay. However, the same convenience has also made it easier for false, altered, and artificially generated material to circulate online. The availability of generative AI has increased this concern because realistic digital content can now be produced with relatively little technical knowledge. Misinformation generally refers to incorrect information that is shared without an intention to mislead, whereas disinformation is created or distributed deliberately to deceive. AI systems can generate fabricated articles, social media messages, synthetic images, modified captions, and artificial audio or video, making it harder to judge the reliability of online information. Deepfake technology represents another major concern in this area. It uses AI and deep-learning techniques to generate or modify faces, voices, expressions, and other parts of digital media. Although deepfakes can have useful applications in entertainment, education, and creative work, they may also be used for impersonation, fraud, reputational harm, and deceptive communication [1], [3], [8]. The detection problem becomes more difficult when different forms of content appear together, such as a misleading statement accompanied by a manipulated image or an AI-generated video. Examining only one component may therefore provide an incomplete view of the content, creating a need for approaches that can consider information from multiple modalities. AI-based detection methods address this problem through Machine Learning (ML), Natural Language Processing (NLP), computer vision, and Deep Learning (DL). Models such as Convolutional Neural Networks (CNNs), transformer architectures, and transfer-learning techniques can learn useful patterns from textual and multimedia data. However, their performance may change when they encounter unfamiliar generation methods, different

languages, compressed media, or data that differs from the training set. Human users may also find highly realistic synthetic media difficult to recognize without technical assistance. For this reason, detection models need to be tested using measures such as accuracy, precision, recall, F1-score, and Receiver Operating Characteristic–Area Under the Curve (ROC-AUC). This research, titled **“Detection of AI-Generated Misinformation and Deepfake Content Using Artificial Intelligence,”** examines machine learning, deep learning, NLP, computer-vision, and multimodal techniques for identifying manipulated and AI-generated content. The study evaluates suitable detection approaches using appropriate datasets and performance measures with the aim of supporting more reliable content verification and improving information integrity in social media environments.



Fig - 1: Overview of AI-Generated Misinformation, Deepfakes, and AI-Based Detection

2. LITERATURE REVIEW

Studies on misinformation detection have gradually moved from basic analysis of textual patterns and source information toward the use of more advanced artificial intelligence techniques. Earlier approaches commonly applied Machine Learning (ML) algorithms such as Support Vector Machines, Logistic Regression, Naïve Bayes, and Random Forest to classify misleading information based on characteristics extracted from available data [2]. The use of Natural Language Processing (NLP) provided additional ways to examine language, context, and relationships between words. Transformer-based models such as BERT further improved contextual text analysis by considering the surrounding information within a sentence [5]. The emergence of Large Language Models (LLMs) has introduced a new difficulty, as automatically generated text can be written in a

style that is similar to human-produced content [6]. This makes detection more dependent on the characteristics of the dataset, language, writing style, and generation system being examined. Research on synthetic media has also developed considerably, particularly in the area of deepfake detection. Datasets and benchmarks such as FaceForensics++ and the DeepFake Detection Challenge have supported the development and comparison of deep-learning-based detection methods [3], [4]. These studies demonstrate the usefulness of learned visual features for identifying manipulated media, while also showing that detector performance can vary across different datasets and manipulation methods.

More recent research has extended detection beyond individual media types by considering combinations of text, images, audio, and video. Such multimodal approaches can provide additional evidence when information from a single source is insufficient for determining whether content is genuine. Research on deepfakes has similarly expanded from image and video manipulation toward audio and other forms of synthetic media [1], [7]. However, several limitations continue to affect practical deployment. Detection models may perform differently when tested on unseen datasets or content generated using newer techniques, while compression, resizing, transcoding, and other social-media processing can remove useful manipulation clues. In addition, combining information from different modalities requires suitable feature-fusion methods and can increase computational complexity. These issues indicate that future detection systems should focus not only on classification accuracy but also on generalization, robustness, multimodal analysis, and practical use in changing social-media environments.

3. AI-BASED DETECTION APPROACHES

Artificial Intelligence has become an important tool for detecting AI-generated misinformation and deepfake content on social media. Recent research has explored a variety of detection techniques that analyse text, images, audio, and video using machine learning and deep-learning methods. This section discusses the major AI-based approaches used for content verification, with a focus on deepfake detection and

multimodal methods designed to improve detection accuracy and reliability.



Fig - 2: AI-Generated Misinformation Research

3.1 DEEPAKE DETECTION USING DEEP LEARNING

Deepfake detection has advanced alongside improvements in AI-based media generation. Early methods focused on identifying visible manipulation artifacts in images and videos, while modern Deep Learning (DL) techniques automatically extract complex visual features from synthetic media. Convolutional Neural Networks (CNNs) have been widely used for image and video analysis, whereas Vision Transformers and hybrid CNN-transformer architectures have improved the ability to capture both local and global patterns. Recent research has further expanded detection to include image, video, audio, and audio-visual analysis, making multimodal deepfake detection an increasingly important area of study [1], [3], [7].

Despite these advances, performance often declines when detection models encounter unfamiliar datasets or new generation techniques. Social media processing, including compression, resizing, cropping, and transcoding, can also remove important forensic clues used by detection systems. These findings demonstrate that reliable deepfake detection requires models capable of adapting to realistic online environments rather than relying only on benchmark datasets [4].

3.2 MULTIMODAL DETECTION, GENERALIZATION, AND ROBUSTNESS

Modern misinformation frequently combines text with images, videos, audio, and user-related information,

making multimodal analysis essential for accurate detection. Researchers have investigated transformer-based models, recurrent architectures, and convolutional methods that integrate information from multiple modalities, often achieving better performance than single-modality systems. However, different media types may contain incomplete or contradictory information, requiring effective fusion strategies to produce reliable predictions [1], [2], [9].

4. RESEARCH GAP AND PROPOSED DIRECTION

The reviewed literature shows that considerable progress has been made in detecting AI-generated misinformation and deepfake content using Machine Learning (ML), Deep Learning (DL), Natural Language Processing (NLP), and multimodal techniques. However, most existing studies focus on a specific type of content, such as text, images, videos, or audio, rather than evaluating multiple content types within a unified detection framework. Many detection models also achieve high performance on benchmark datasets but experience reduced accuracy when applied to unfamiliar datasets, new generative models, or real-world social media content affected by compression, resizing, and other platform-related modifications [1], [3], [4].

Another limitation identified in the literature is the difficulty of adapting detection systems to rapidly evolving generative AI technologies. The characteristics used to identify AI-generated content can change as newer generation models become more advanced, making previously trained detectors less effective. In addition, challenges related to multimodal feature fusion, explainability, computational efficiency, and cross-dataset generalization remain unresolved, indicating that reliable detection requires more robust and adaptable approaches [1], [3], [4].

To address these limitations, the present study investigates AI-based techniques capable of analysing both misinformation and deepfake content using suitable machine learning, deep learning, NLP, and computer-vision methods. The research emphasizes evaluating detection approaches using appropriate datasets and standard performance metrics such as accuracy, precision, recall, F1-score, and ROC-AUC. By considering multiple forms of digital content and current detection

challenges, the study aims to contribute toward more effective and practical AI-driven content verification for social media environments [1], [2], [3].

5. PROBLEM STATEMENT

The rapid advancement of generative Artificial Intelligence (AI) has made it easier to create realistic text, images, audio, and video, increasing the spread of AI-generated misinformation and deepfake content on social media. A single post may combine misleading text with manipulated images, synthetic audio, or deepfake videos, making authenticity difficult to determine. Although existing detection methods have achieved promising results, many are designed for specific media types and often struggle with unseen datasets, evolving generation techniques, media compression, and real-world social media conditions [1], [2], [3], [4].



Fig - 3: Challenges of AI-Generated Misinformation and Deepfake Content

Another challenge is that AI-generated content is not always misinformation, while misinformation can also be created by humans. Therefore, an effective detection system should distinguish between synthetic content and misleading information rather than treating them as the same problem. Considering these challenges, this study investigates AI-based methods capable of detecting manipulated and AI-generated content using textual and multimedia characteristics. The research also evaluates suitable datasets, preprocessing methods, detection models, and performance metrics to develop a more reliable approach for social media content verification [1], [2].

6. OBJECTIVES OF THE STUDY

The primary objective of this study is to investigate the application of Artificial Intelligence (AI) techniques for detecting AI-generated misinformation and deepfake content and to evaluate their effectiveness in identifying manipulated or misleading digital content.

The specific objectives of the study are as follows:

1. To examine the characteristics of AI-generated misinformation and deepfake content on social media.
2. To investigate existing Artificial Intelligence, Machine Learning (ML), Deep Learning (DL), and Natural Language Processing (NLP) techniques used for detection.
3. To identify suitable datasets and preprocessing methods for analysing textual and multimedia content.
4. To implement an AI-based detection approach capable of classifying authentic and manipulated content.
5. To evaluate model performance using accuracy, precision, recall, F1-score, and ROC-AUC.
6. To compare different detection approaches and identify their strengths and limitations.
7. To examine challenges related to evolving AI-generated content, unseen generation techniques, and real-world social media environments.
8. To propose recommendations for improving AI-based misinformation and deepfake detection.

7. HYPOTHESIS OF THE STUDY

This study investigates whether Artificial Intelligence (AI)-based techniques can effectively detect AI-generated misinformation and deepfake content. Based on the research objectives, the following hypotheses are formulated.

7.1 NULL HYPOTHESIS (H_0)

There is no statistically significant difference in the detection performance of AI-based techniques when identifying authentic content and AI-generated or manipulated misinformation and deepfake content.

7.2 ALTERNATIVE HYPOTHESIS (H₁)

There is a statistically significant difference in the detection performance of AI-based techniques when identifying authentic content and AI-generated or manipulated misinformation and deepfake content.

7.3 MODEL PERFORMANCE HYPOTHESIS

H₀₂: Advanced AI and deep-learning techniques do not significantly improve detection performance compared with conventional machine-learning approaches.

H₁₂: Advanced AI and deep-learning techniques significantly improve detection performance compared with conventional machine-learning approaches.

The hypotheses will be evaluated using experimental results obtained from selected datasets and AI-based detection models. Performance will be assessed through accuracy, precision, recall, F1-score, and ROC-AUC, and the final interpretation will be based on the comparative performance of the evaluated models.

8. METHODOLOGY

This study follows a systematic methodology for evaluating AI-based techniques used to detect AI-generated misinformation and deepfake content. The methodology includes dataset selection, data preprocessing, model implementation, and performance evaluation to ensure a reliable comparison of different detection approaches.



Fig - 4: Workflow of the Proposed Methodology

8.1 DATASET SELECTION

Suitable publicly available datasets containing authentic and AI-generated content will be selected for experimentation. The datasets may include text-based misinformation, manipulated images, deepfake videos, and synthetic audio, depending on availability and research scope.

8.2 DATA PREPROCESSING

Before model training, the collected data will undergo preprocessing to improve quality and consistency. Text data will be cleaned through tokenization, normalization, and stop-

word removal, while image and video data will be resized, normalized, and formatted for model input.

8.3 MODEL IMPLEMENTATION

The study will examine both conventional and advanced AI techniques. Machine Learning algorithms such as Support Vector Machines (SVM), Logistic Regression, and Random Forest will be considered alongside Deep Learning models, transformer-based architectures, Natural Language Processing models, and computer-vision techniques for multimodal detection.

8.4 PERFORMANCE EVALUATION

The implemented models will be evaluated using standard classification metrics, including accuracy, precision, recall, F1-score, and Receiver Operating Characteristic–Area Under the Curve (ROC-AUC). These measures will be used to compare model effectiveness and identify the strengths and limitations of different AI-based detection approaches.

9. CONCLUSIONS

This study focuses on investigating Artificial Intelligence-based techniques for detecting AI-generated misinformation and deepfake content in social media environments. By examining existing detection methods, identifying research gaps, and evaluating suitable AI models using standard performance metrics, the research aims to contribute to the development of more reliable automated content verification systems. The findings are expected to support future improvements in digital media security and help address the evolving challenges posed by generative AI technologies.

ACKNOWLEDGEMENT

I sincerely express my gratitude to Prof. Aarti Sanjay Gawai, Assistant Professor, Department of Information Technology and Computer Science, D.G. Ruparel College of Arts, Science and Commerce, Mumbai, for her valuable guidance, encouragement, and continuous support throughout this research. I also thank D.G. Ruparel College of Arts, Science and Commerce for providing the academic

environment and resources that contributed to the successful completion of this study.

REFERENCES

- [1] Mirsky, Y., Lee, W.: The Creation and Detection of Deepfakes: A Survey. *ACM Computing Surveys* 54(1) (2021) 1–41.
- [2] Shu, K., Sliva, A., Wang, S., Tang, J., Liu, H.: Fake News Detection on Social Media: A Data Mining Perspective. *ACM SIGKDD Explorations* 19(1) (2017) 22–36.
- [3] Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: FaceForensics++: Learning to Detect Manipulated Facial Images. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (2019) 1–11.
- [4] Dolhansky, B., Howes, R., Pflaum, B., Baram, N., Canton Ferrer, C.: The DeepFake Detection Challenge Dataset. *arXiv preprint arXiv:2006.07397* (2020).
- [5] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: *Proceedings of NAACL-HLT* (2019) 4171–4186.
- [6] Brown, T., Mann, B., Ryder, N., et al.: Language Models are Few-Shot Learners. In: *Advances in Neural Information Processing Systems (NeurIPS)* 33 (2020) 1877–1901.
- [7] Agarwal, S., Farid, H., Gu, Y., He, M., Nagano, K., Li, H.: Protecting World Leaders Against Deep Fakes. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2019).
- [8] Westerlund, M.: The Emergence of Deepfake Technology: A Review. *Technology Innovation Management Review* 9(11) (2019) 39–52.
- [9] Vaccari, C., Chadwick, A.: Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception and Uncertainty. *Social Media + Society* 6(1) (2020).

BIOGRAPHIES



Mr. Shubham Jaidas

Bhise

M. Sc. Information
Technology

Student, Dept. of IT/CS,

D. G. Ruparel College
Ms. Aarti Sanjay Gawai

M. Sc. Information
Technology

Asst. Prof., Dept. of
IT/CS,

D. G. Ruparel College

