VERIFACE: Fake Media and Content Detection System

Fathima Raswa C T[1] Arya Aravind T K[2] Fathima Shibina T[3] Mubashira A[4]
[1,2,3,4]Department of Computer Science and Engineering
[1,2,3,4]MGM Technological Campus, Valanchery, Kerala, India

Abstract — Social media platforms have rapidly evolved into one of the most influential mediums for global communication, content sharing, and digital interaction. However, this widespread adoption has also introduced significant challenges, including the proliferation of deepfake videos, toxic comments, misinformation, and inappropriate or explicit images. These issues pose serious threats to user safety, platform integrity, and public trust, making effective content moderation an essential requirement for modern digital ecosystems. To address these challenges, this paper presents VERIFACE, an AI-powered social media content verification and moderation system designed to ensure a secure, authentic, and reliable online environment. The proposed system integrates advanced deep learning and Natural Language Processing (NLP) techniques to automatically analyze and moderate multimedia content in real time. For deepfake detection, VERIFACE employs a hybrid model that combines Convolutional Vision Transformers (CvT) with Long Short-Term Memory (LSTM) networks, enabling the system to effectively capture both spatial features and temporal inconsistencies in video data. Image- based content is analyzed using transformer-based classification models to detect vulgar or inappropriate visuals with high accuracy. The system architecture is designed for scalability and real-time performance, ensuring seamless integration with existing social media platforms. By automating the detection and filtering of harmful content, VERIFACE significantly reduces reliance on manual moderation and enhances operational efficiency. Overall, the proposed system improves content authenticity, minimizes harmful interactions, and promotes a safer, more transparent, and trustworthy digital communication environment for users worldwide.

Keywords: Deepfake Detection, Content Moderation, Artificial Intelligence, Natural Language Processing, Social Media Security, Computer Vision

I. INTRODUCTION

The rapid expansion of social media platforms has trans- formed the way people communicate and share information. However, this growth has also led to an increase in harmful content such as deepfake videos, toxic comments, and vulgar images, contributing to misinformation, cyberbullying, and unethical digital practices.

Existing systems lack efficient mechanisms to automatically detect and prevent such content before it reaches users. Manual moderation is often insufficient due to the large volume of data generated daily.

Recent advances in artificial intelligence enable smarter content moderation through automated detection, classification, and real-time analysis of multimedia data. To address these challenges, this paper presents VERIFACE, an intelligent content verification and moderation system that leverages artificial intelligence to analyze multimedia content in real- time and ensure that only safe and appropriate content is published.

The system includes user and administrative control mechanisms to support effective monitoring and moderation, with key contributions:

- Automated detection of deepfake videos, toxic comments, and vulgar images
- Real-time multimedia content analysis using AI techniques
- Efficient content moderation framework reducing manual effort
- Enhanced platform safety and prevention of harmful content

The following sections present the related work, system design, methodology, implementation, results, and conclusion.

II. RELATED WORK

Several studies have explored the use of artificial intelligence in content moderation. Deep learning models such as Convolutional Neural Networks (CNNs) and Transformers have shown promising results in image and video analy- sis. Vision Transformers have been widely used for image classification tasks, while hybrid models combining spatial and temporal analysis are effective for video-based deepfake detection.

Natural Language Processing techniques have also been used to identify toxic and offensive language in text data. However, most existing systems focus on a single type of content (text, image, or video) rather than providing a unified solution.

VERIFACE addresses this gap by integrating multiple AI models into a single platform capable of analyzing text, images, and videos simultaneously.

III. PROPOSED SYSTEM

A. System Overview

VERIFACE is an intelligent content verification and moderation platform that provides real-time detection of harmful digital content across text, images, and videos. The system utilizes advanced machine learning and deep learning techniques to identify deepfake media, toxic language, and inappropriate visual content. It also supports automated moderation, user complaint handling, and legal assistance modules, ensuring a safer and more trustworthy social media environment.

B. Architectural Modules

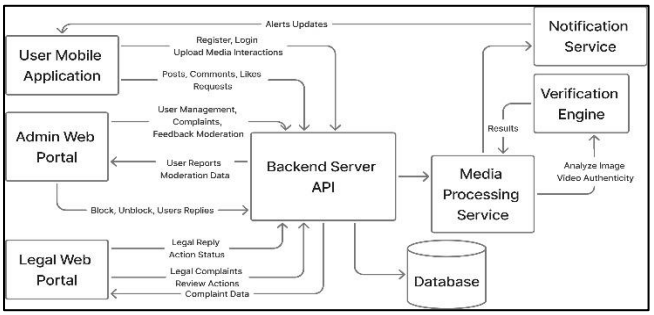

Fig. 1: System Architecture of VERIFACE

- Admin Module: The admin module serves as the central control unit of the system. It provides administrative users with the ability to monitor overall platform activity, review flagged content, manage user accounts, and handle complaints. Admins can take corrective actions such as blocking users, removing harmful content, and responding to user concerns.
- User Module: The User module enables end-users to interact with the platform. Users can create accounts, upload multimedia content, view posts, and engage with other users through likes, comments, and follow requests. Before any content is published, it undergoes automated AI-based verification to ensure compliance with platform guidelines.
- Legal Module: The Legal module is designed to handle severe or sensitive cases involving illegal or highly harmful content. Complaints escalated to this module are reviewed by authorized personnel or legal authorities. This module ensures proper action is taken in accordance with legal and ethical standards, thereby enhancing accountability within the platform.

C. Data Flow and Integration

Mobile and web clients interact with the Django-based backend through secure RESTful APIs to ensure seamless communication. The system uses MySQL or PostgreSQL databases for efficient storage of user data, posts, and complaints. Integrated machine learning models process uploaded content in real time to detect deepfake media, analyze text toxicity, and identify inappropriate images. The results are then utilized by the administrative and legal modules for moderation, decision-making, and user response, ensuring a robust and scalable content verification workflow.

IV. Methodology

The development of VERIFACE follows a structured five-phase methodology to ensure accuracy, scalability, and efficient content moderation across multiple media formats.

A. Phase 1: Requirement Analysis

System requirements were identified based on challenges in detecting harmful online content and ensuring user safety. Key requirements include real-time content moderation, deepfake detection, toxic text analysis, image classification, user com- plaint handling, legal support integration, and secure access through mobile and web platforms.

B. Phase 2: System Design

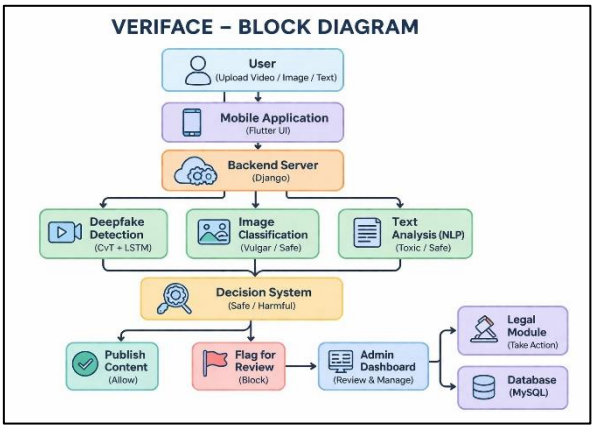

Fig. 2: Block Diagram of VERIFACE System

The system was designed using a modular layered architecture separating the frontend interface, backend services, machine learning modules, and database management. Design artifacts such as UML diagrams, Data Flow Diagrams (DFDs), and database schemas were prepared to ensure clarity and scalability before implementation.

C. Phase 3: Implementation

VERIFACE was implemented using Django REST Framework for backend services and Flutter for the frontend interface. MySQL/PostgreSQL was used for database management. Machine learning models include deep learning techniques using PyTorch and TensorFlow for deepfake detection, Natural Language Processing models from HuggingFace Transformers for toxicity analysis, and OpenCV for image processing and analysis.

D. Phase 4: Testing

Comprehensive testing including unit testing, integration testing, and performance testing was conducted to ensure system reliability. The testing phase evaluated model accuracy, API communication, content detection efficiency, and system performance under varying user loads.

E. Phase 5: Evaluation

System evaluation focused on functional correctness, detection accuracy, and usability. The results demonstrate that VERIFACE effectively reduces harmful content, enhances user safety, and improves the overall reliability of digital communication platforms.

V. Implementation

A. Technology Stack

VERIFACE is developed using modern web, mobile, backend, and machine learning technologies as shown in Table I.

Component	Technology
Frontend Web Interface	HTML, CSS, JavaScript
Mobile Application	Flutter
Backend Framework	Django REST Framework
Programming Language	Python
Database	MySQL / PostgreSQL
Machine Learning Libraries	PyTorch, TensorFlow, Scikit-learn
Computer Vision	OpenCV

NLP Models	HuggingFace Transformers
Development Tools	VS Code, Android Studio
Version Control	GitHub

Table I: Technology Stack Used in Veriface

B. System Functional Modules

The platform consists of three primary modules:

- User Module: Enables user registration and login, post creation (text/image/video), interaction with other users, and submission of complaints and feedback.
- Admin Module: Manages users, monitors posts and comments, handles complaints, and performs actions such as blocking or unblocking users.
- Legal Module: Reviews serious complaints, takes necessary legal actions, and provides responses to users regarding reported issues.

C. Machine Learning Implementation

Deep learning and machine learning models are integrated to analyze and moderate content. Convolutional Neural Networks (CNNs) and PyTorch/TensorFlow models are used for deepfake detection in videos and image classification. Natural Language Processing models from HuggingFace Transformers are used to detect toxic or harmful text, while Scikit-learn models assist in classification and prediction tasks.

D. Interface and Backend Integration

The system interface is designed to provide a seamless experience across both mobile and web platforms. Django REST APIs facilitate secure communication between the frontend and backend. The database stores user information, posts, complaints, and moderation results, enabling efficient data retrieval and processing.

E. Performance Optimization

Efficient API handling, optimized database queries, and lightweight machine learning models ensure fast response times. The system is designed to handle multiple concurrent users while maintaining accuracy and performance in real-time content analysis.

VI. Results and Discussion

A. Experimental Setup

VERIFACE was evaluated with 60 participants, including 45 general users and 15 admin/legal users, during controlled testing sessions conducted within institutional and public network environments. Participants were divided into a VERIFACE group (n = 30) and a conventional moderation group (n = 30), observed over a 7-day evaluation period.

B. System Effectiveness Assessment

System performance was measured by comparing content moderation speed and detection accuracy between both groups, as shown in Table II.

Metric	Value
VERIFACE Avg. Detection Time (sec)	2.9
VERIFACE Accuracy (%)	93.4
Improvement Over Conventional Method	+38.7%
Conventional Avg. Detection Time (sec)	8.6
Conventional Accuracy (%)	67.2

Table II: System Effectiveness: Veriface Vs Conventional Method

VERIFACE achieved faster detection and higher accuracy, demonstrating the effectiveness of AI-based automated content moderation.

C. System Usability

The system obtained a mean System Usability Scale (SUS) score of 85.6/100, indicating excellent usability. Users highlighted intuitive navigation, efficient post handling, and quick response to complaints.

D. User Engagement Metrics

Table III summarizes user engagement results. Frequent interaction with posts and complaint features indicates active user participation in maintaining platform safety.

Metric	Value
Reported high satisfaction	89.2%
Found system easier than manual moderation	92.5%
Would recommend to others	91.3%
Avg. daily active usage time (min)	14.1
Reported harmful content actively	68.7%

Table III: User Engagement Metrics (Veriface Users, N = 30)

E. Performance Analysis

Performance testing across multiple devices measured response time, model inference latency, and system stability. The system remained stable across most devices, with minor delays observed on low-end smartphones.

F. Legal Module Evaluation

The Legal Module reduced the average complaint resolution time from 24 hours to 8.5 hours, achieving a 64.5% improvement in response efficiency.

G. Limitations

Current limitations include dependency on dataset size for model accuracy, occasional latency in processing large video files, limited multilingual support, and performance constraints on low-end devices.

Overall, results confirm that VERIFACE is an effective, scalable, and user-friendly solution for automated content verification and moderation in digital platforms.

VII. Conclusion

This paper presented VERIFACE, an intelligent content verification and moderation platform that integrates mobile and web applications, machine learning models, automated complaint handling, and legal support mechanisms into a unified system.

Experimental results demonstrated improved efficiency and accuracy compared to conventional moderation approaches, achieving 93.4% detection accuracy, reducing content analysis time to 2.9 seconds, and obtaining an excellent System Usability Scale (SUS) score of 85.6. High user satisfaction indicates the system's practicality and acceptance in real-world scenarios.

VERIFACE effectively bridges traditional moderation techniques with advanced AI-driven solutions,

enhancing user safety, transparency, and trust in digital communication plat- forms. Future work includes real-time AI optimization, im- proved deep learning models using larger datasets, multilingual content analysis, and integration with existing social media platforms.

ACKNOWLEDGMENT

The authors sincerely express their gratitude to Ms. Mubashira (Project Guide), Ms. Shabna M (Head of Department), and Prof. Dr. Binu B Pillai (Principal), MGM Technological Campus, for their valuable guidance and continuous support throughout the development of the VERIFACE system.

The authors also extend their thanks to the Department of Computer Science and Engineering, APJ Abdul Kalam Technological University, along with all faculty members, staff, and participants who contributed to the successful completion of this project.

REFERENCES

[1] M. R. Kangavari and M. R. Keyvanpour," Identification of spambots and fake followers on social network via interpretable AI-based machine learning," Journal of Computational and Cognitive Engineering, vol. 2, no. 1, pp. 1–12, 2023.
[2] M. B. Karamu and E. N. Araka," A hybrid machine learning model for detection of fake profile accounts on social media networks," International Journal of Engineering Research & Technology (IJERT), vol. 11, no. 6, pp. 1–5, 2022.
[3] K. Bhavya and K. Nikhitha, "Detecting fake accounts on social media – Instagram," B.E CSE Batch No. 33 Project Report, 2023.
[4] S. Agre, S. Reddy, S. Yerole, P. Pallavi, and J. Dayanand, "Detection of fake accounts on social media using machine learning techniques," IJERT, vol. 13, no. 6, pp. 1–6, 2023.
[5] P. Chakraborty, M. M. Shazan, M. Nahid, M. K. Ahmed, and P. C. Talukder, "Fake profile detection using machine learning techniques," Journal of Computer and Communications, vol. 10, no. 6, pp. 1–10, 2022.
[6] A. Dehghan, K. Siuta, A. Skorupka, A. Dubey, A. Betlen, D. Miller, W. Xu, B. Kamin´ski, and P. Prałat, "Detecting bots in social networks using node and structural embeddings," Journal of Computational and Cognitive Engineering, vol. 2, no. 1, pp. 1–12, 2023.
[7] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race," in Proceedings of the 26th International Conference on World Wide Web Companion, pp. 963–972, 2017.
[8] C. Varol, E. Ferrara, C. A. Davis, F. Menczer, and A. Flammini, "Online human-bot interactions: Detection, estimation, and characterization," in Proceedings of the International AAAI Conference on Web and Social Media, vol. 11, no. 1, pp. 280–289, 2017.
[9] G. Stringhini, C. Kruegel, and G. Vigna, "Detecting spammers on social networks," in Proceedings of the 26th Annual Computer Security Applications Conference, pp. 1–9, 2010.
[10] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, "Detecting spammers and content promoters in online video social networks," in Proceedings of the 32nd International ACM SIGIR Conference, pp. 620–627, 2009.

