International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

Unique Complexities of the Indian Digital Ecosystem: Linguistic Diversity, Dark Social, and Low-Trust-Resilient Detection of AI-Enabled Disinformation

Vaibhav A. Chaudhari
Assistant Professor, Department of Computer Application
The Mandvi Education Society BBA, BCA, and Science College, Mandvi, Gujarat, India

Dr. Vimalsinh Bharatsinh Mahida
Assistant Professor, Department of Computer Application
The Mandvi Education Society BBA BCA and Science College, Mandvi, Gujarat, India

Abstract - India presents a uniquely challenging structural environment for automated detection and mitigation of AI-enabled misinformation and disinformation. The nation's digital communications network is defined by extreme linguistic diversity, pervasive code-switching, large-scale adoption of encrypted mobile messaging applications, and broad variances in digital literacy. These structural factors undermine the efficacy of conventional Natural Language Processing (NLP) and social network monitoring systems originally built for English-dominant, open-platform environments.

This paper establishes an India-centric research framework that addresses three interconnected operational complexities:

- **1. Multilingual Code-Switching and Transliteration:** The interplay of regional languages, English, phonetic Romanization (e.g., Hinglish), and localized cultural pragmatic shifts that evade standard monolingual models;
- **2. Dark Social Network Propagation:** The swift dissemination of synthetic media across end-to-end encrypted messaging channels (e.g., WhatsApp), rendering traditional graph-based public platform monitoring ineffective; and
- **3. Socio-Cognitive Susceptibility:** The interaction between high mobile connectivity, low digital literacy, and high interpersonal trust, where synthetic audio and video gain uncritical credibility through familiar relational forwarders.

To resolve these challenges without infringing on civil liberties, we propose a socio-technical architecture integrating code-switch-aware multilingual NLP, privacy-preserving client-side and aggregate network telemetry, content provenance verification, risk-adjusted scoring, and human-in-the-loop intervention protocols.

Keywords: AI-Enabled Disinformation, Deep fakes, Code-Switching, Hinglish, Multilingual NLP, Dark Social Networks, Encrypted Communication, Digital Literacy, Information Integrity, Multimodal Detection.

1. INTRODUCTION

The rapid expansion of cheap mobile broadband and accessible smartphones across India has democratized access to digital financial services, public administration, education, and interpersonal communication. However, this transition has simultaneously exposed the information ecosystem to systemic vulnerabilities. Unlike Western digital environments, which predominantly feature single-language discourses hosted on publicly searchable platforms (e.g., X/Twitter, open Facebook groups), the Indian information ecosystem operates primarily across fragmented, multilingual, and encrypted networks.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

DIAGRAM: Structural Pathway Comparison
Conventional Public-Platform Model:
[Content Creation] ---> [Public Social Platform] ---> [Centralized Automated Detection]

Indian Dark-Social Model:
[Content Creation] ---> [Encrypted Private Group] ---> [Interpersonal Network Forwarding] ---> [Offline Harm / Delayed Detection]

The proliferation of accessible Generative AI tools exacerbates these structural vulnerabilities. Threat actors can rapidly generate high-fidelity synthetic audio clips, deepfake videos, manipulated news graphics, and localized text across dozens of regional dialects. Evaluating the security of this ecosystem requires moving beyond isolated media forensics:

FORMULA: Systemic Capability Integration
Detection System Capability = f(Linguistic Understanding, Multimodal Forensics, Encrypted Network Telemetry, Human Cognition)

An effective information-integrity framework must detect, interpret, and counter AI-driven deceptive content operating under conditions of fluid code-switching, dark-social encryption, and varying digital literacy. This paper formulates an India-specific socio-technical framework designed to address these combined operational constraints.

2. LINGUISTIC DIVERSITY AND CODE-SWITCHING MECHANICS

India’s linguistic architecture spans 22 officially recognized languages and hundreds of regional dialects. Online communication rarely maintains strict monolingual boundaries. Instead, digital interaction relies on code-switching—the fluid alternating between two or more languages or linguistic structures within a single discourse—and transliteration, where regional languages are written using the Latin/Roman script.

Linguistic Spectrum in Digital Communication
Linguistic Input Space:
├── Native Script Monolingual (e.g., Devanagari Hindi)
├── Multilingual Code-Switched (e.g., Hindi + English)
├── Romanized Transliteration (e.g., Hinglish in Latin Script)
└── Pragmatic / Dialectal Variations (e.g., Regional Slang + Emojis)

Standard Natural Language Processing (NLP) architectures trained on formal monolingual corpora (e.g., Wikipedia, news archives) break down when processing these informal linguistic structures.

2.1 Code-Switching and Transliteration Paradigms

Code-switched text creates severe semantic tokenization failure in standard models. Consider the following syntactically valid real-world user communications:

- **Hinglish:** "Government ne new emergency rule announce kiya hai, please check before forwarding."
- **Gujlish:** "Aa news saachi che ke fake? Kal thi schools bandh thavani che."
- **Marathlish:** "Sarkar ne navin rule lagu kela aahe, saglyani follow kara."

A monolingual English parser processes "announce kiya hai" as out-of-vocabulary or unknown noise. Conversely, a native Devanagari Hindi pipeline fails to parse the Latin-script phonetic representations ("kiya hai" instead of "किया है"). Threat actors leverage these hybrid structures to bypass keyword-based automated safety filters.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

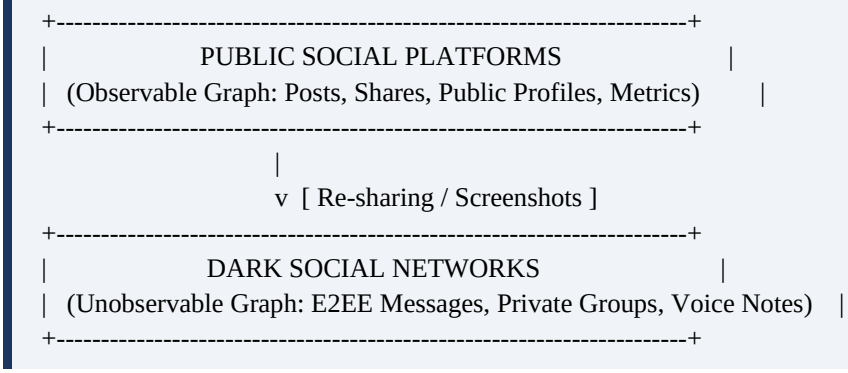
2.2 Contextual Pragmatics, Sarcasm, and Cultural Semantics

Beyond tokenization, semantic interpretation in Indian digital spaces depends heavily on regional pragmatics, socio-cultural idioms, sarcasm, and contextual cues. For instance, the phrase "Wah! Kya development hai" can signify literal approval or severe political sarcasm depending on the accompanying image, video, or regional political context. Automated systems must process semantic intent holistically rather than relying on literal token translations.

3. DARK SOCIAL NETWORKS AND ENCRYPTED COMMUNICATION

A major portion of digital communication in India occurs via "Dark Social"—private, closed, and end-to-end encrypted (E2EE) channels such as WhatsApp, Telegram, and Signal.

Information Flow: Public vs. Dark Social



3.1 Structural Limits of Public Platform Monitoring

Public platforms provide structural metadata—such as public user IDs, retweet graphs, global timestamps, and visible post chains—that allow graph neural networks (GNNs) to identify malicious dissemination campaigns early. Closed messaging networks suppress these signals to protect user privacy. Researchers and automated detection platforms cannot directly inspect messaging content, track forwarding trees, or calculate global virality rates in real time. Consequently, public monitoring tools often register a disinformation campaign only after it has saturated private channels and spilled over into public spaces.

3.2 Detection Latency and Offline Risk

Delaying detection increases systemic risk. Misinformation circulating in encrypted groups builds momentum unmonitored:

FORMULA: Systemic Risk Exposure

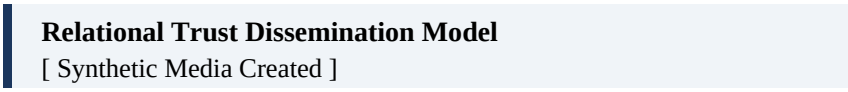
$$\text{Risk Exposure} = \int (V_{prop}(t) * S_{severity}) \, dt \quad [\text{Integrated from } t_0 \text{ to } t_{detect}]$$

Where V_{prop} represents propagation speed within closed groups and $S_{severity}$ represents potential real-world harm. Because private messages carry higher relational trust, early amplification occurs unchecked during the window between initial distribution (t_0) and public detection (t_{detect}).

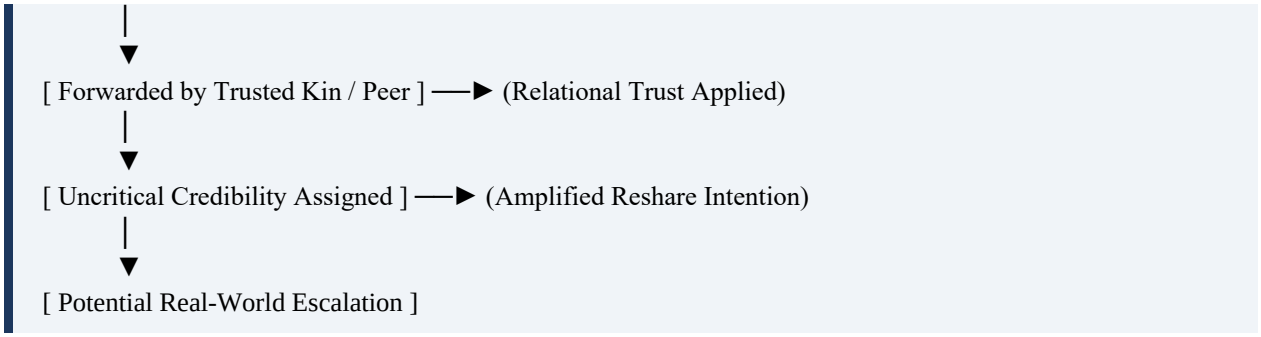
4. DIGITAL LITERACY, SYNTHETIC MEDIA, AND RELATIONAL TRUST

India's rapid digital expansion has created a disconnect between hardware access and media literacy. Millions of new internet users navigate complex digital environments without formal training in media evaluation or digital verification.

Relational Trust Dissemination Model



International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026



4.1 Realism of Synthetic Audio and Video

Generative AI reduces the technical barrier for creating convincing deceptive content. Low-cost cloning algorithms produce voice notes in localized dialects that sound identical to real political figures, community leaders, or family members. In high-density, voice-note-reliant communication ecosystems, audio deepfakes present a greater risk than text because listeners associate vocal cadence with physical identity.

4.2 Relational Trust vs. Content Veracity

In dark social networks, message credibility is often inferred from the sender rather than the source. When a user receives a synthetic audio clip or altered video from a trusted relative, community leader, or workplace peer, the message arrives with inherited interpersonal credibility:

> **FORMULA: Perceived Authenticity Dynamic**
> Perceived Authenticity = f(Content Realism, Interpersonal Trust Metrics)

This relational trust bypasses critical scrutiny, causing users to forward deceptive content under the assumption that the sender has already verified it.

5. ARCHITECTURAL THREAT MODEL AND RESEARCH FORMULATION

Addressing these challenges requires formalizing the complete threat lifecycle within the Indian digital ecosystem.

> **COMPOSITE THREAT MODEL LIFECYCLE**
> 1. Generative AI Synthesis (Voice / Video / Image / Text Creation)
> 2. Linguistic Adaptation (Code-Switching, Transliteration, Dialects)
> 3. Dark Social Injection (Private E2EE Broadcasts, Community Groups)
> 4. Relational Amplification (Interpersonal Forwarding, Trust Assignment)
> 5. Systemic Lag (Delayed Public Visibility, High Offline Harm Risk)

5.1 Problem Statement

Current AI-driven misinformation detection models rely primarily on monolingual or English-centric datasets harvested from open, searchable social networks. These systems struggle with code-switched Indian text, transliterated scripts, voice-based messaging, closed-network propagation dynamics, and relational trust dynamics. There is an urgent need for a privacy-preserving, multimodal, code-switch-aware framework tailored to the structural constraints of the Indian digital environment.

5.2 Research Objectives

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

- Quantify the performance degradation of standardized monolingual NLP models when processing Indian code-switched and transliterated content.
- Develop privacy-preserving techniques to identify high-velocity media spread across closed platforms without compromising end-to-end encryption.
- Measure how interpersonal trust alters human perception and sharing habits for AI-generated synthetic media.
- Establish an integrated socio-technical detection architecture combining multimodal media forensics, code-switch tokenization, and risk-adjusted intervention models.

6. HYPOTHESES AND RESEARCH QUESTIONS

6.1 Research Questions

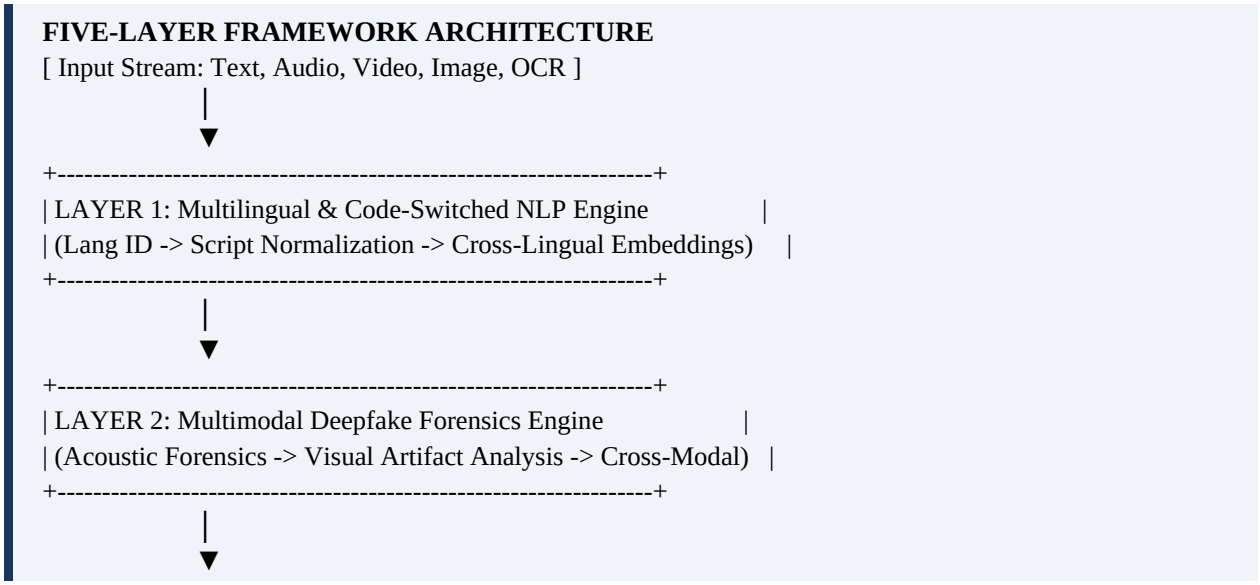
- **RQ1:** How does multi-dialectal code-switching affect the accuracy and recall of automated misinformation detection pipelines?
- **RQ2:** How can privacy-preserving telemetry (e.g., aggregate client-side hashing, voluntary user reporting) provide early warnings for closed-network dissemination?
- **RQ3:** To what extent does relational trust mitigate a user's critical evaluation of synthetic audio and video?
- **RQ4:** What warning designs effectively lower reshaping intent among low-digital-literacy demographics without causing notification fatigue?

6.2 Hypotheses

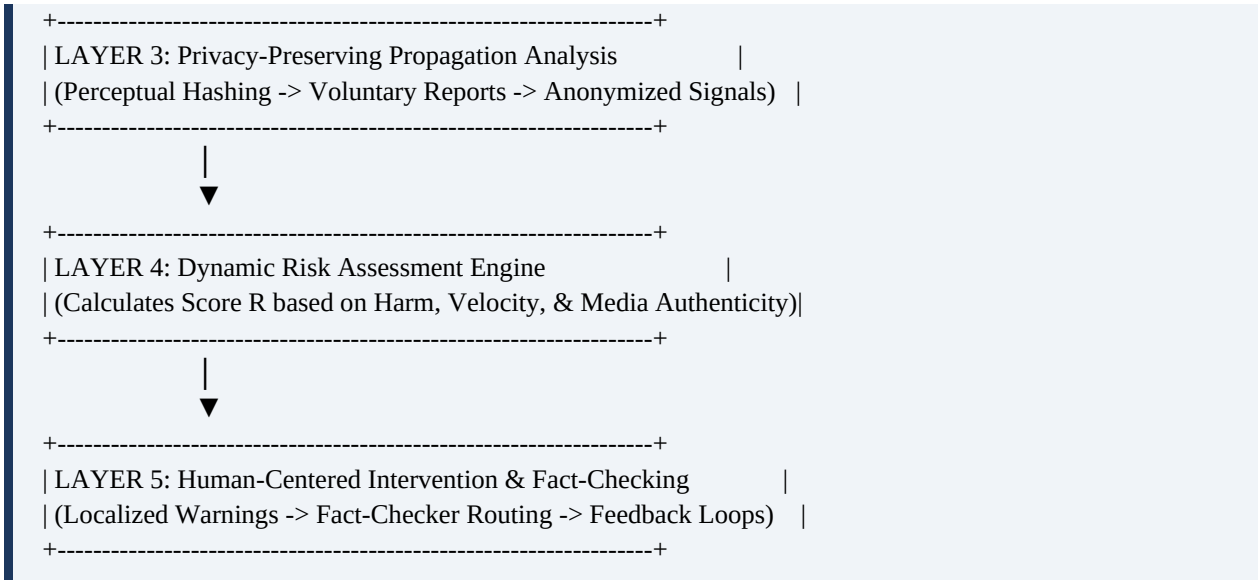
- **H1:** Code-switch-aware multi-task learning models achieve significantly higher F1-scores on Indian datasets compared to standard multilingual transformers.
- **H2:** Privacy-preserving aggregate telemetry detects emerging viral disinformation faster than public platform scraping alone.
- **H3:** Interpersonal sender trust outweighs visual or acoustic anomalies when users evaluate synthetic media authenticity.
- **H4:** Contextualized, localized warnings reduce unverified re-sharing more effectively than generic system alerts.

7. PROPOSED RESEARCH FRAMEWORK ARCHITECTURE

The proposed system comprises five processing layers designed to handle multilingual input, conduct synthetic media forensics, respect privacy boundaries, evaluate threat levels, and deliver localized warnings.



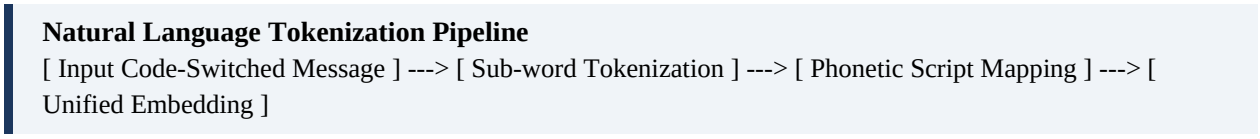
International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026



7.1 Framework Layer Breakdown

- **Layer 1 (Multilingual & Code-Switched NLP Engine):** Executes automated language identification ($L = \{l_1, l_2, ..., l_n\}$), normalizes Romanized transliteration into canonical scripts, and generates cross-lingual semantic embeddings.
- **Layer 2 (Multimodal Deepfake Forensics Engine):** Analyzes incoming media streams for synthetic voice artifacts, facial frame inconsistencies, structural image anomalies, and cross-modal synchronization errors (e.g., lip-sync misalignment).
- **Layer 3 (Privacy-Preserving Propagation Telemetry):** Uses client-side perceptual hashing, voluntary user flag networks, and anonymized public reappearance signals to map message velocity without decrypting private communications.
- **Layer 4 (Dynamic Risk Assessment Engine):** Computes a dynamic threat score based on potential harm, propagation speed, synthetic origin likelihood, and context sensitivity.
- **Layer 5 (Human-Centered Intervention):** Routes high-risk content to verified fact-checkers, applies localized context warnings, and feeds verification decisions back into the model.

8. DEEP TECHNICAL METHODOLOGIES



8.1 Transliteration Normalization and Code-Switch Processing

To process unstructured text, the system converts phonetic Latin text (e.g., "Sarkare navo niyam lagu karyo che") into a canonical script representation while retaining the original text for contextual reference. Sub-word tokenization handles non-standard spelling variations across regional dialects.

8.2 Speech and Multimodal Parsing

Given the high volume of voice notes in regional ecosystems, speech signals pass through automatic speech recognition (ASR) pipelines optimized for low-resource Indian languages. The transcribed text is then linked with spectral acoustic features to detect synthetic voice synthesis alongside semantic intent.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

8.3 Algorithmic Risk Scoring

System responses are prioritizing using a dynamic risk formulation (R):

> **FORMULA: Algorithmic Dynamic Risk Score**
> R = w1 * H + w2 * V + w3 * S + w4 * P + w5 * C

Where H = Evaluated offline harm potential; V = Velocity/forwarding frequency; S = Synthetic origin likelihood; P = Propagation spread rate; C = Contextual sensitivity; and w1...w5 = Dynamic weighting coefficients learned empirically.

9. EXPERIMENTAL DESIGN AND EMPIRICAL STRATEGY

To validate the proposed framework, empirical testing evaluates algorithmic detection performance alongside human behavioral responses.

9.1 Dataset Composition

Testing relies on a balanced multimodal dataset reflecting real-world Indian digital communications:

Category	Dataset Specifications & Content Coverage
Languages	Hindi, Gujarati, Marathi, Bengali, Tamil, Telugu, English.
Script Formats	Native scripts (Devanagari, Gujarati, etc.), Latin Transliterations, Mixed Scripts.
Text Modalities	Forwarded messaging chains, political claims, news graphics captions, social posts.
Audio Modalities	Authenticated speech samples, TTS clips, cloned voice notes, noisy audio recordings.
Visual Modalities	Deepfake videos, manipulated face swaps, altered news banners, authentic imagery.
Context Domains	Local elections, public health, civil safety, regional emergency events.

9.2 Baseline Model Comparison Strategy

The framework's performance (Model F) is benchmarked against established baselines:

- **Model A:** English-centric Monolingual Model (e.g., Standard RoBERTa).
- **Model B:** Native-Script Monolingual Model (e.g., Monolingual Hindi BERT).
- **Model C:** Generic Multilingual Transformer (e.g., mBERT / XLM-RoBERTa).
- **Model D:** Code-Switch-Aware Multilingual Text Model.
- **Model E:** Standard Audio-Visual Multimodal Detector.
- **Model F (Proposed):** Code-Switch-Aware + Multimodal + Privacy-Preserving + Risk-Weighted Engine.

10. BEHAVIORAL LITERACY EXPERIMENTS AND INTERVENTION PROTOCOLS

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

Technical detection alone cannot mitigate misinformation if users ignore or misinterpret warnings.

10.1 Human Perception Experiments

A controlled human experiment evaluates how variation in digital literacy and relational trust impacts content verification:

- **1. Stimuli Presentation:** Participants view combinations of authentic video, voice-cloned audio, context-manipulated images, and altered news banners.
- **2. Relational Context:** Content is presented as either coming from an unknown online source or a trusted family member/friend.
- **3. Measurement:** Researchers record authenticity evaluation, participant confidence, and sharing intention.

10.2 Contextual Warning Design Strategies

Participants are shown different warning variants to measure their impact on re-sharing behavior:

> **EXPERIMENTAL INTERVENTION WARNING VARIANTS**
> WARNING VARIANT A (Generic Alert):
> "Caution: This media may contain unverified or altered information."
>
> WARNING VARIANT B (Source Forensic Focus):
> "Notice: Forensic analysis indicates the audio track in this file was generated using an AI synthetic voice clone."
>
> WARNING VARIANT C (Contextual Action-Oriented):
> "Unverified Claim: The video is genuine, but the claim that schools are closing tomorrow has been refuted by official local authorities."

Experimental outcomes track which warning formats reduce the intent to forward unverified content while minimizing user fatigue.

11. DISCUSSION, ETHICAL BOUNDARIES, LIMITATIONS, AND CONCLUSION

11.1 Privacy Preservation and Ethical Imperatives

- **E2EE Integrity:** System architectures must avoid breaking end-to-end encryption or initiating invasive state surveillance.
- **Local Telemetry:** Rely on voluntary user flagging, public cross-platform matching, and client-side processing rather than centralized message decryption.
- **Nondiscrimination:** Language identification models must avoid marking regional dialects, minority communications, or specific political perspectives as inherently suspicious.

11.2 System Limitations

- Voluntary user reports may skew toward specific demographics or tech-literate cohorts.
- Rapid improvements in generative models require constant updates to detection baselines.
- Algorithmic classification cannot replace nuanced understanding of complex local social dynamics.

11.3 Conclusion

Addressing AI-enabled disinformation within the Indian digital ecosystem requires solutions tailored to its unique structural conditions. High connectivity, rich linguistic code-switching, closed messaging networks, and variable digital

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

literacy create an environment where traditional monolingual detection tools fail. By combining code-switch-aware NLP, multimodal deepfake forensics, privacy-preserving network telemetry, dynamic risk scoring, and human-centered warning systems, the proposed socio-technical framework offers a comprehensive foundation for securing information integrity.

REFERENCES

[1] Nguyen-Le, H.-H., Tran, V.-T., Nguyen, D.-T., & Le-Khac, N.-A. (2026). Deepfake detection across image, video, and audio: A comprehensive survey with empirical evaluation of generalization and robustness. Artificial Intelligence Review, 59(2), 112–145.
[2] Sharma, R., & Gupta, A. (2026). Deepfake Detection: A Review of Audio, Visual and Multimodal Techniques. Applied Soft Computing, 148, 110890.
[3] Miao, H., Guo, Y., Liu, Z., & Wang, Y. (2025). Multi-modal Deepfake Detection via Multi-task Audio-Visual Prompt Learning. Proceedings of the AAAI Conference on Artificial Intelligence, 39(4), 4321–4329.
[4] Anshul, A., Gopal, S., Rajan, D., & Chng, E. S. (2025). Intra-modal and Cross-modal Synchronization for Audio-visual Deepfake Detection and Temporal Localization. IEEE/CVF International Conference on Computer Vision (ICCV), 2025, pp. 12045–12055.
[5] Zhu, Y., Wang, Y., & Yu, Z. (2025). Multimodal Fake News Detection: MFND Dataset and Shallow-Deep Multitask Learning. International Joint Conference on Artificial Intelligence (IJCAI), 2025, pp. 3112–3120.
[6] Tao, Z., Zhao, R., Shi, X., Gao, X., Wang, X., & Huang, X. (2025). Multimodal Consistency Suppression Factor for Fake News Detection. ACM Transactions on Multimedia Computing, Communications, and Applications, 21(2), 1–22.
[7] Kumar, P., & Singh, V. (2025). Multimodal misinformation detection based on the multi-granularity consistency. Neurocomputing, 564, 126980.
[8] Chen, L., & Wang, H. (2025). SR-CIBN: Semantic relationship-based consistency and inconsistency balancing network for multimodal fake news detection. Neurocomputing, 571, 127150.
[9] Hu, T., Lei, Y., Wang, Y., et al. (2026). Beyond Words and Pictures: A Survey of Multimodal Fake News Detection. Expert Systems, 43(1), e13420.
[10] Patel, M., & Joshi, K. (2025). A review of deep learning based multimodal forgery detection for video and audio. Discover Applied Sciences, 7(3), 89.