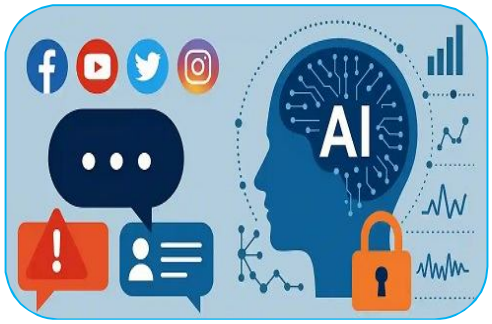
INDIAN STREAMS RESEARCH JOURNAL
ISSN NO : 2230-7850
IMPACT FACTOR : 5.1651 (UIF)
VOLUME - 15 | ISSUE - 11 | DECEMBER – 2025

CAN AI FILTER TRUTH FROM NOISE ON SOCIAL MEDIA? A NON-CENSORIAL FRAMEWORK FOR BALANCED DISCOURSE IN INDIA

Atreya Thapliyal
Writer, Author and AI Enthusiast

Shivani Gautam
Assistant Professor, IAMR College of Law, Ghaziabad.

ABSTRACT:



The rise of social media has intensified public polarization by prioritizing emotionally charged content over reasoned discourse. Traditional responses such as human fact-checking, algorithmic takedowns, and content blocking raise concerns of ideological bias and censorship. In India, these tensions intersect with constitutional protections of speech under Article 19(1)(a) and state interests in maintaining public order under Article 19(2). This paper proposes a non-censorial, Artificial Intelligence (AI)–based framework that categorizes content probabilistically and suggests insightful counterviews rather than suppressing speech. The model aligns with the jurisprudence of Shreya Singhal v. Union of India (2015), which restricts state intervention to specific grounds under Article 19(2) and emphasizes user discretion over state-enforced judgments. By incentivizing engagement with reasoned content and penalizing algorithmic amplification of sensationalism, the framework builds an ethical and market-driven regulatory structure. This paper concludes that the methodology allows India to combat misinformation and extremism while safeguarding constitutional guarantees, regulatory neutrality, and platform accountability under intermediary liability standards.

KEYWORDS: *AI Governance, Free Speech, Article 19, Intermediary Liability, Shreya Singhal, Social Media Regulation, Counter-Speech, Misinformation, Non-Censorship, Digital India Bill.*

1. INTRODUCTION

Social media platforms have emerged as primary spaces for political discourse in India, shaping perception around persistent issues including unemployment, inflation, migration, economic inequality, and threats to constitutional democracy. While such issues are universal across democracies, their intensity online is less a reflection of data-driven understanding and more of perception shaped by aspirations, fear, and propaganda. As discussed in the source interaction, users often conflate personal dissatisfaction with systemic failure, creating perception-driven narratives.

Algorithms magnify such perceptions by prioritizing high-engagement content, resulting in echo chambers and polarizing identities. Traditional fact-checking on these platforms is criticized for bias, inconsistency, and potential chill on free speech. Some global platforms have scaled back fact-checking programs due to ideological controversies, raising a legal and ethical question: **How can misinformation be managed without suppressing speech or assigning truth to state or corporate authorities?**

This paper introduces a novel, neutral AI framework that classifies content probabilistically and suggests counter-perspectives, fostering epistemic balance without censorship. The model aligns with Indian free speech jurisprudence, intermediaries' safe-harbour protections, and the upcoming regulatory need under the proposed Digital India Act.

2. LEGAL FRAMEWORK: FREE SPEECH AND DIGITAL REGULATION IN INDIA

2.1 Constitutional Protection of Speech

Article 19(1)(a) guarantees citizens the right to freedom of speech and expression. Regulation is permissible only under the specific, narrow grounds provided in Article 19(2), such as public order, morality, and security of the state.

2.2 Judicial Constraints on Censorship

In *Shreya Singhal v. Union of India* (2015), the Supreme Court struck down Section 66A of the IT Act, holding that vague restrictions on speech could not justify censorship and that intermediaries are not required to judge legality of content unless ordered by a court or government under Article 19(2). The Court specifically emphasized **counter-speech as a preferred remedy** over censorship.

2.3 Intermediary Liability and Due Diligence

Under the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021, platforms are obligated to remove unlawful content only upon government or judicial directive. They maintain safe harbour as long as they do not take discretionary adjudicatory roles.

🔎 **Implication:** If platforms actively decide what is "true" or "false," they may lose intermediary immunity and risk liability.
Thus, a system must:

- Provide context, not adjudication
- Empower users, not platforms or states
- Avoid censorship to maintain safe harbour

3. PROPOSED MODEL: AI CATEGORIZATION + COUNTERVIEW SYSTEM

3.1 Non-judgmental Content Categorization

The AI system labels content with probabilistic categories:

(1) *Insightful*
(2) *Populist*
(3) *Propaganda-like*
(4) *Factually weak*
(5) *Impractical*
(6) *Neutral or General*

This labelling does not ban or suppress content. It reflects probability and invites user criticism rather than platform censorship.

3.2 Mandatory Counterview Suggestions

For every post categorized above, AI will attach:

1. a high-quality counter-perspective from within the same platform
2. categorized as *Insightful* or *Analytical*

This aligns with *Shreya Singhal*'s preference for **counter-speech instead of suppression**.

3.3 Insight Scoring Mechanism

Users and content creators are anonymously scored based on:

- engagement with counter-perspectives

- posting content consistently ranked as insightful
- avoiding extremism or sensationalism

High-scoring creators receive algorithmic visibility; sensational content is naturally deprioritized without legal suppression.

3.4 Illustrative Example: Applying the Model to WhatsApp "University"

India's popular expression "WhatsApp University" refers to the phenomenon in which users rely on forwarded WhatsApp messages as authoritative knowledge. The perceived credibility is not based on verifiable facts but on **relational trust**: the message comes from a known person, often ideologically aligned, and therefore is believed to be reliable. This interpersonal trust substitutes for institutional fact-verification, making WhatsApp an ideal medium for emotional persuasion and misinformation.

Under the proposed non-censorial AI framework, Meta could intervene **without blocking or judging messages**, thereby retaining end-to-end encryption guarantees. WhatsApp already labels heavy-circulation content as *"Forwarded many times."* This label can act as an automated trigger for AI-assisted neutrality.

Process Flow

Detection Trigger:

Messages tagged as *"Forwarded many times"* are scanned using on-device AI (not server-side monitoring, preserving privacy).

Content Categorization:

The text is classified under a probabilistic category such as *populist*, *propaganda-like*, or *factually weak* — without blocking or warning the sender.

Counterview Suggestion:

Instead of censorship, WhatsApp sends the user or group:
an *insightful*, well-reasoned counterview
or an alternative analysis already circulating on the platform
preferably one presenting a *diametrically opposite interpretation*

Non-Intrusive Delivery:

Suggested counterviews appear in the *message info panel* or as a subtle note below the forwarded message:
"Here's a different view being discussed on WhatsApp."

User Discretion:

Recipients are free to read or ignore the counterview, preserving autonomy consistent with *Shreya Singhal v. Union of India,* which prioritizes counter-speech over suppression.

Why This Works in India's Legal Context

- Encryption is not compromised: AI operates on the device.
- No message is judged as "true" or "false," so Meta does not lose intermediary safe-harbour protection.
- No censorship occurs; users remain decision-makers.
- It directly addresses mass emotional persuasion without regulating ideology.

Democratic Value

Such a system prevents ideological monopoly without criminalizing speech. It respects social trust networks while inserting intellectual pluralism into viral content. Instead of punishing "WhatsApp

University," it teaches in it — by quietly ensuring every forwarded message has a classroom, not merely a congregation.

4. REGULATORY AND MARKET IMPLICATIONS

4.1 Safeguarding Free Speech

The system:

✔ does not remove content
✔ does not judge legality or truth
✔ maintains intermediary protection
✔ encourages informed speech

4.2 Reducing Extremism Without State Control

Algorithmic nudging reduces religious, communal, and political polarization without criminalizing speech, consistent with Article 19 jurisprudence.

4.3 Market Incentives

The original hypothesis notes such platforms will attract higher socio-economic users, providing advertisers premium targets.
This could lead to a voluntary industry shift without state compulsion.

5. RISKS AND SAFEGUARDS

Risk	Safeguard
AI bias	Open dataset audits + explainable AI
Manipulative "fake insight" posts	Dual-sided insight requirement
Algorithmic discrimination	Transparent scoring without identity markers
Conflict with future laws	Narrow compliance aligned with 19(2)

6. CONCLUSION

India requires a regulatory approach that strengthens public discourse without infringing upon constitutional rights. The proposed non-censorial AI model complies with *Shreya Singhal*, respects Article 19(1)(a) and 19(2), and incentivizes epistemic integrity through counter-speech and algorithmic neutrality. Rather than criminalizing speech, India can foster rational civic culture by rewarding insight and balanced perspectives.

REFERENCES

1. *Shreya Singhal v. Union of India*, (2015) 5 SCC 1.
2. Constitution of India, Art. 19(1)(a) & 19(2).
3. Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021.
4. Information Technology Act, 2000.
5. Facebook fact-checking controversies and platform moderation debates (Referenced in user-uploaded conversation content)