International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 11 | Nov - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Content Moderation and Freedom of Expression in Indian Cyberspace

Author[1]: Naman Kumar, Department of Journalism and Mass Communication, Tecnia Institute of Advanced Studies, (Affiliated to GGSIP University, Delhi)

Author [2]: Dr Ashish Kumar, Associate Professor, Department of Journalism and Mass Communication, Tecnia Institute of Advanced Studies, (Affiliated to GGSIP University, Delhi)

Abstract

In recent years, the Indian cyberspace has witnessed a rapid expansion of digital communication, social media platforms and user-generated content. Alongside these opportunities for expression, there is increasing concern about how content moderation practices—by platforms, intermediaries and the state—impact the fundamental right of freedom of expression in India. This paper explores the evolving legal and policy framework for online content moderation in India, examines the tensions between moderation and free speech, reviews existing literature on the topic, and identifies key challenges and recommendations for balancing the twin imperatives of democratic expression and responsible digital governance. The study finds that while content moderation is necessary to address harmful or illegal online speech, there is a risk of over-moderation, lack of transparency, and chilling effects on legitimate expression. It concludes that a multi-stakeholder approach, clearer rules, greater transparency and due process are needed to maintain freedom of expression in Indian cyberspace.

Keywords: Digital Censorship, Platform Governance, Online Speech Regulation, Internet Freedom in India and Algorithmic Moderation.

Introduction

The advent of the Internet and social media has transformed how individuals express themselves, ac cess information and engage with each other in society. In India, the democratization of content creation and sharing has empowered millions of users to voice opinions, mobilise for causes, and participate in public discourse. At the same time, the growth of user-generated content has posed new challenges: harassment, hate speech, misinformation, extremist propaganda, and other forms of harmful or unlawful speech. In response, both private platforms and the Indian state have introduced content moderation mechanisms, regulations and governance frameworks aimed at curbing harmful online speech.

However, the process of content moderation—deciding what expression is permitted, what is restricted, and what is removed—raises profound questions about the protection of the fundamental right to freedom of expression under Article 19(1)(a) of the Constitution of India, as well as its reasonable restrictions under Article 19(2). The interplay between platform responsibility, intermediary liability, state regulation and individual rights is complex, and in the Indian context features additional dimensions such as linguistic diversity, socio-cultural plurality, digital divide, and evolving regulatory practice.

This paper seeks to map the contours of content moderation in Indian cyberspace, analyse how freedom of expression is being impacted, review relevant academic and legal literature, and derive an understanding of the key issues and possible way-forward. The paper is structured as follows: first, an overview of the regulatory framework and context; second, a literature review of existing research; third, an analytical discussion of how content moderation and free speech interact in India; and finally conclusions and recommendations.

Literature Review

The literature on content moderation and freedom of expression is growing rapidly—both globally and in the Indian context. Some of the salient themes and findings from existing studies are outlined below.

1. Global perspectives on content moderation & free speech

Recent works address how platforms moderate content across jurisdictions, and how this impacts free speech. For example, the volume *Free Speech in the Puzzle of Content Regulation: Insights from the West and the Global South* provides a comparative study of regulatory models (self-regulation, co-regulation, external regulation) and highlights the tensions between platform moderation and free expression.

Another survey by Kiritchenko et al. reviews automated abuse-detection systems in online platforms and notes that while they help reduce harmful content, they often risk silencing under-represented groups, raising challenges for fairness and non-discrimination.
Thus, globally there is acknowledgement that the scale, speed and complexity of moderation (especially algorithmic moderation) threatens, unless carefully managed, the right to expression.

2. Indian regulatory and legal context

In the Indian context, the landmark judgment Shreya Singhal v. Union of India (2015) is foundational: the Supreme Court struck down Section 66A of the Information Technology Act, 2000 for being disproportionate and in violation of the freedom of speech guarantee.
Further, the 2021 intermediary and digital media ethics rules — the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021 — impose specific obligations on social media intermediaries, including disclosure of grievance officers, traceability of originators of messages (in certain cases) and mandated takedowns of "unlawful" content.
Some critical literature emphasises that these rules, while aimed at accountability, raise serious concerns about over-broad definitions, lack of safeguards, and potential chilling effect on expression. For example, a study of online harassment by the Software Freedom Law Center (Indian chapter) concluded that while moderation and legal remedies are needed, they must not morph into censorship.
Another piece reviews the mechanism of content regulation on social networking websites in the Indian context—including the balancing act between freedom of expression and content regulation.
Also, a recent article in the *Indian Journal of Legal Review* titled *"A Critical Analysis on the Impact of Social Media Platforms Content Moderation Policies on Freedom of Speech with Special Reference to Chennai"* examines how platform moderation policies and user experiences intersect with freedom of speech in Indian cities.
Thus, Indian scholarship is increasingly engaging with the intersection of law, platform governance and user rights in the digital space.

3. Key themes emerging from literature

From the above and additional sources, we can summarise the key themes:

- **Scale and speed of content**: The volume of user-generated content is enormous and moderation must work at scale, often relying on automated tools. This raises risks of error, bias and unfair suppression of legitimate speech.
- **Platform versus state regulation**: Platforms are private entities yet often play quasi-public roles because they host public discourse. Handling moderation raises questions of private power, transparency, accountability and redress mechanisms.

- **Intermediary liability and safe-harbour**: The Indian law on intermediaries (e.g., Section 79 IT Act) and the rules for takedowns raise questions about when platforms lose safe-harbour, and how that impacts their moderation decisions.
- **Chilling effects and over-moderation**: If moderation mechanisms are vague, arbitrary or opaque, they risk deterring users from legitimate expression for fear of removal or sanction.
- **Cultural, linguistic and contextual diversity**: In India, multiple languages, socio-cultural dynamics and access disparities complicate moderation practices and raise the risk of marginalising voices from under-represented groups.
- **Transparency, accountability and due process**: Literature emphasises the need for published transparency reports, user redress mechanisms, human oversight (not purely algorithmic), and meaningful consultation in rule-making.
- **Emerging technologies (AI/moderation tools)**: The integration of AI in moderation brings benefits (scale) but also challenges of fairness, bias, error, and lack of human nuance.

In sum, the literature reveals a rich but contested field, where the imperative for content moderation (to prevent harm) and the imperative for freedom of expression (to enable democratic discourse) must be balanced delicately.

Analytical Discussion: Content Moderation vs Freedom of Expression in Indian Cyberspace

In this section, we analyse how the legal, institutional and technological frameworks in India mediate the interplay between content moderation and the freedom of expression, noting key issues and gaps.

The Legal Framework

In India, freedom of speech and expression is guaranteed under Article 19(1)(a) of the Constitution of India. However, this right is subject to "reasonable restrictions" under Article 19(2) in the interests of sovereignty and integrity of India, security of the State, public order, decency or morality, etc.
In cyberspace, the main statute is the Information Technology Act, 2000, and its rules and amendments including the Intermediary Guidelines and Digital Media Ethics Code Rules, 2021. Under Section 79 of the IT Act, intermediaries (platforms) enjoy safe-harbour protection from liability for third-party content, provided they comply with due diligence and take down content upon actual knowledge or court order. The Shreya Singhal case (2015) read-down Section 79 and struck down Section 66A of the IT Act as unconstitutional for being vague and disproportionate. The 2021 Rules impose additional obligations: platforms must appoint grievance officers; major intermediaries must publish policies and deploy tools; traceability of "first originator" is required in certain cases; takedown obligations within specified timeframes.

While this framework recognises the need for moderation and draws a broad structure, several issues arise.

Key Issues and Tensions

1. Vague definitions and broad discretion: What constitutes "unlawful content", "public order", "harm", "offensive" is often vague. The rules allow takedowns on broad grounds. Some studies warn that such ambiguity leads to arbitrary removals and undermines freedom of speech.
2. Over-moderation and chilling of expression: Platforms, to avoid liability, may over-comply and remove borderline content, leading to self-censorship by users. In India this risk is heightened due to the fear of sanctions, account suspensions, or legal risks. The SFLC report flagged that online harassment and threats act as forms of censorship.
3. Platform power and transparency deficits: The major platforms (social media intermediaries) wield significant power to decide content removals. Yet the literature points out a lack of transparency: inadequate disclosure of removal reasons, inadequate redress mechanisms, limited human review. The risk is that private decisions shape public discourse

without adequate checks.

4. Traceability, privacy & encryption issues: The 2021 Rules' requirement for traceability of the first originator of messages (in some cases) creates tension with encryption and privacy rights. It may undermine user anonymity and thereby hamper free expression, particularly for dissenting or marginalized voices

5. Technological/AI-driven moderation vs human nuance: While scale demands automated tools, literature cautions that AI moderation lacks human nuance, contextual understanding of Indian languages, dialects, culture, humour, satire. This can lead to wrongful removals of legitimate speech. The global survey by Kiritchenko et al. cautions that under-represented groups can be disproportionately silenced.

6. Socio-cultural and linguistic diversity in India: India's linguistic plurality (22 official languages, hundreds of dialects) complicates moderation accuracy. The dominant platform moderation tools may be calibrated more for major languages, leaving minority languages more vulnerable to errors. Also cultural sensitivity and local context matter: what is acceptable in one region may be offensive in another. The literature emphasises the need for culturally aware moderation.

7. State regulation vs platform self-regulation: A key tension exists between state regulatory intervention (e.g., takedown orders, blocking of sites, government notifications) and platform self-governance (community guidelines, internal moderation policies). The Indian ecosystem requires both, but the question is how to ensure that state regulation doesn't become excessive censorship and that platform moderation doesn't become arbitrary. The IT Rules 2021 attempt to combine both obligations, but critics argue that the balance is off.

8. Due-process and redress mechanisms: Freedom of expression implies that if speech is moderated/removed, there should be fair process, clarity of reason, opportunity to challenge. The literature indicates that Indian frameworks lack robust procedural safeguards in many cases. This lacuna increases the risk of legitimate expression being silenced without recourse.

Impacts on Freedom of Expression

Given the above issues, the moderation practices and regulatory mechanisms in India impact freedom of expression in multiple ways:

- The fear of removal or penalty may lead to self-censorship, especially among dissenting voices, minority language users, or marginalised communities.
- Legitimate public interest speech (critique of government, activism, satire) may be chilled if policies are vague and enforcement opaque.
- Over-moderation may narrow the diversity of voices in Indian cyberspace, undermining the democratic marketplace of ideas.
- Differential access to moderation resources: large, mainstream language users may have more nuanced moderation, while regional languages and smaller platforms may suffer inconsistent treatment.
- The trust in platforms and regulatory processes is impacted: users may feel they don't know why content was removed, how decisions were taken. Lack of transparency undermines confidence in moderation fairness.
- On the flip side, inadequate moderation can allow harmful content (hate speech, misinformation, incitement to violence) which itself undermines the right of others to express, and can distort public discourse. Thus, moderation is necessary—but must be balanced.

Pathways and Recommendations

Based on the literature and analysis, some suggestions to better align content moderation with freedom of expression in Indian cyberspace include:

- **Clear definitions and guidelines**: Regulatory rules and platform policies should clearly define unlawful/harmful content, with examples and thresholds, to reduce arbitrariness.

- **Transparency and accountability**: Platforms should publish detailed transparency reports (number of takedowns, reasons, languages, redress outcomes). Government takedown/block orders should be made public where feasible (consistent with security/public-order).
- **Fair redress and due process**: Users must be informed of reasons for removal, given opportunity to appeal, and have access to independent review (internal + external).
- **Human-in-the-loop moderation and linguistic/cultural diversity**: Moderation systems should combine automated tools with human moderators trained in local languages and contexts. Investment is needed in minority-language moderation.
- **Proportionate regulation and safe-harbour clarity**: Intermediary liability rules should maintain a strong safe-harbour for platforms so they are not incentivised to over-remove, and the government's role in mandates must respect constitutional safeguards.
- **Multi-stakeholder governance**: Platform companies, government, civil society, academic researchers and user communities should engage collaboratively in rule-making, monitoring and evaluation. Pre-legislative consultation is key.
- **Digital literacy and user empowerment**: Users should be educated about moderation policies, recourse processes, privacy rights, and how to contest removals or raise grievance.
- **Periodic review and empirical research**: The impact of moderation rules in India should be studied empirically (language-wise, platform-wise, region-wise) to identify biases, unintended effects and refine regulation accordingly.
- **Safeguards for freedom of expression**: The regulation of online content must align with constitutional jurisprudence linking to Article 19(1)(a) and (2), ensuring that restrictions are lawful, necessary, proportionate, and subject to meaningful scrutiny.

Conclusion

In this rapidly evolving digital environment, content moderation and freedom of expression are engaged in a delicate dance. On one hand, moderation is essential to maintain safe, inclusive online spaces and to prevent harms such as hate speech, harassment, misinformation and incitement. On the other hand, freedom of expression is a foundational democratic right, vital for debate, dissent, creativity and the formation of public opinion. In the Indian context, with its diverse languages, cultures and socio-digital landscapes, the challenge is particularly acute.

The legal and regulatory framework in India — including the IT Act, intermediary liability rules, and platform obligations — sets out the architecture for moderation but leaves important gaps: in definitions, transparency, redress, due process and context sensitivity. The literature indicates that while some progress has been made, risks remain of over-moderation, chilling of speech, uneven treatment of languages and communities, and private power without sufficient accountability. For India to navigate this terrain effectively, a balanced, calibrated approach is needed—one that empowers users, engages platforms responsibly, ensures state regulation is constitutionally anchored and transparent, and respects the rich pluralism of India's cyberspace. As digital expression continues to expand and evolve, so too must the ecosystems of moderation, governance and rights protection adapt.

References

- Kiritchenko, S., Nejadgholi, I., & Fraser, K. C. "Confronting Abusive Language Online: A Survey from the Ethical and Human Rights Perspective." *arXiv*, 2020.
- "Free Speech in the Puzzle of Content Regulation: Insights from the West and the Global South." Soorya Balendra (Ed.). Springer, 2024.
- "The Complex Land of Cyber Laws and Freedom of Speech in India: The Role of Social Media." LegalServiceIndia.
- "Impact of Cyber Laws on Social Media." LegalServiceIndia.
- "A Critical Analysis on the Impact of Social Media Platforms Content Moderation Policies on Freedom of Speech with Special Reference to Chennai." Lisa S. & Hanushka Srinivasan. *Indian Journal of Legal Review* (IJLR), 2024.
- "Online Harassment: A Form of Censorship." Press Release, Software Freedom Law Center India.
- "Navigating the Privacy-Freedom Dilemma: The Impact of AI on Content Moderation and Free Speech." Khalid Ali & Sandeep Arthur Kumar.
- "Online Content Regulation on Social Networking Website." Ashok Kumar & … (2018).
- Shivendu Kumar Rai. Globalization and digital violence against women in new media. Int J Appl Res 2017;3(6):961-966.
- Riya Bansal, Dr.Shivendu Kumar Rai. (2025). How AI Enhances Creativity in Content Writing. In IJSRED-International Journal of Scientific Research and Engineering Development (Vol. 8, Number 3, pp. 639–645). Zenodo. https://doi.org/10.5281/zenodo.15408007.
- Dr. Shivendu Kumar Rai,Ms. Priyanka Singh, "EFFECTIVENESS AND CHALLENGES OF ONLINE LEARNING: A CASE STUDY ON STUDENTS OF HIGHER EDUCATION", International Journal of Creative Research Thoughts (IJCRT), ISSN:2320-2882, Volume.10, Issue 9, pp.a240-a249, September 2022, Available at :http://www.ijcrt.org/papers/IJCRT2209036.pdf