INDIAN JOURNAL OF LEGAL REVIEW

VOLUME 5 AND ISSUE 9 OF 2025

INSTITUTE OF LEGAL EDUCATION

INDIAN JOURNAL OF LEGAL REVIEW

APIS – 3920 – 0001 | ISSN - 2583-2344

(Open Access Journal)

Journal's Home Page – https://ijlr.iledu.in/

Journal's Editorial Page - https://ijlr.iledu.in/editorial-board/

Volume 5 and Issue 9 of 2025 (Access Full Issue on - https://ijlr.iledu.in/volume-5-and-issue-10-of-2025/)

Publisher

Prasanna S,

Chairman of Institute of Legal Education

No. 08, Arul Nagar, Seera Thoppu,

Maudhanda Kurichi, Srirangam,

Tiruchirappalli – 620102

Phone : +91 94896 71437 - info@iledu.in / Chairman@iledu.in

INDIAN JOURNAL OF LEGAL REVIEW [IJLR – IF SCORE – 7.58]

VOLUME 5 AND ISSUE 10 OF 2025

APIS – 3920 - 0001 *(and)* ISSN - 2583-2344

Published by

Institute of Legal Education

https://iledu.in

EMPOWERING MARGINALIZED VOICES: BALANCING HATE SPEECH REGULATION WITH FREE EXPRESSION FOR VULNERABLE COMMUNITIES IN INDIA'S DIGITAL ECOSYSTEM

AUTHOR - SHOBHIT BANSAL, DEPARTMENT OF LAW, CHISHTI LANGUAGE UNIVERSITY, LUCKNOW, INDIA.
EMAIL- SHOBHITEXAMLKO@GMAIL.COM

BEST CITATION - SHOBHIT BANSAL, EMPOWERING MARGINALIZED VOICES: BALANCING HATE SPEECH REGULATION WITH FREE EXPRESSION FOR VULNERABLE COMMUNITIES IN INDIA'S DIGITAL ECOSYSTEM, *INDIAN JOURNAL OF LEGAL REVIEW (IJLR),* 5 (10) OF 2025, PG. 506-512, APIS – 3920 – 0001 & ISSN - 2583-2344

Abstract

This paper examines the complex intersection of hate speech regulation and free expression in India's expanding digital ecosystem, with particular focus on marginalized communities. Through analysis of existing regulatory frameworks, legal precedents, and socio-cultural contexts, this research illuminates the challenges faced by vulnerable populations in exercising their right to expression while being protected from harmful content. The study employs a qualitative approach, examining case studies and policy implementations to evaluate their effectiveness. Findings suggest that current regulatory mechanisms often fail to adequately protect marginalized voices while simultaneously limiting legitimate expression from these communities. The paper proposes a balanced framework that centers vulnerable populations in policy development, advocates for contextual understanding of hate speech, and emphasizes community participation in content moderation processes. This research contributes to ongoing discourse on digital rights in India by highlighting the need for nuanced approaches that both combat hate speech and preserve free expression for those most vulnerable to silencing.

Keywords: Digital Rights, Hate Speech Regulation, Free Expression, Marginalized Communities, India, Content Moderation

1. Introduction

India's digital landscape has undergone remarkable transformation in recent years, with over 760 million internet users as of 2023[1]. This digital expansion has created unprecedented opportunities for expression, particularly for historically marginalized communities who have gained platforms to articulate their experiences, challenges, and aspirations. However, this digital democratization has coincided with increasing incidents of targeted hate speech and discriminatory content online[2]. This tension—between protecting vulnerable communities from harmful speech while ensuring their right to expression—represents one of the most significant challenges in India's evolving digital rights framework.

The complexity of this issue is amplified by India's diverse socio-cultural landscape, marked by intersecting identities based on caste, religion, gender, sexuality, language, ethnicity, and disability. Each of these identity markers carries historical contexts of discrimination that manifest differently in digital spaces[3]. Traditional approaches to hate speech regulation often fail to account for these nuances, resulting in frameworks that either inadequately protect marginalized communities or inadvertently restrict their legitimate expression.

This paper investigates this critical tension, exploring how hate speech regulations impact marginalized communities' ability to participate in digital discourse. It examines current

regulatory frameworks, their implementation, and their effects on vulnerable populations, while proposing approaches that better balance protection with expression. The central research question asks: How can India develop regulatory frameworks that effectively protect marginalized communities from online hate speech while preserving their right to free expression in digital spaces?

2. Theoretical Framework and Literature Review

2.1 Conceptualizing Hate Speech in the Indian Context

Hate speech exists on a spectrum of harmful expression that includes disinformation, incitement, harassment, and threats. In the Indian context, scholars like Bhatia (2021) have argued that hate speech must be understood through the lens of historical power dynamics and structural inequalities[4]. Traditional Western liberal frameworks of free speech often fail to account for India's unique socio-cultural context, where speech acts cannot be separated from their potential to reinforce existing hierarchies and discrimination.

According to Narrain (2016), hate speech in India must be conceptualized not merely as offensive expression but as speech that "contributes to a climate that normalizes discrimination against vulnerable groups"[5]. This perspective recognizes that seemingly isolated instances of hateful content online can collectively create environments hostile to marginalized communities' participation.

2.2 Digital Rights and Marginalized Communities

The literature on digital rights has increasingly recognized the differentiated impacts of online regulation on marginalized communities. Kovacs and Ranganathan (2019) note that vulnerable populations often experience what they term "double silencing"—facing both targeted harassment that discourages their participation and overly broad content moderation that removes their legitimate expression[6]. This perspective challenges the notion that there exists a simple trade-off between protection and expression.

Research by the Internet Democracy Project (2021) documents how women from Dalit communities face intersectional forms of abuse online that combine caste-based and gendered harassment[7]. Similarly, studies by Gupta (2020) demonstrate how religious minorities in India experience coordinated campaigns of harassment that aim to exclude them from digital public spheres[8]. These findings highlight the need for approaches to regulation that recognize the varied experiences of different marginalized groups.

2.3 Regulatory Approaches and Their Limitations

India's approach to regulating online content has evolved through multiple legal frameworks, including provisions of the Indian Penal Code, the Information Technology Act, and more recently, the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021. Scholars like Abraham and Hickok (2018) have critiqued these frameworks for their ambiguity and potential for overbroad interpretation[9].

Chima and Kaur (2022) argue that India's regulatory approach often prioritizes rapid content removal without sufficient attention to context, leading to disproportionate impacts on already marginalized voices[10]. Their analysis suggests that content moderation systems—both state-mandated and platform-operated—frequently lack the cultural competency to distinguish between harmful speech and legitimate political expression from marginalized communities.

This literature review reveals significant gaps in understanding how regulatory frameworks specifically affect marginalized communities' expressive capabilities, highlighting the need for research that centers these communities' experiences in evaluating and developing regulatory approaches.

INDIAN JOURNAL OF LEGAL REVIEW [IJLR – IF SCORE – 7.58]
VOLUME 5 AND ISSUE 10 OF 2025
APIS – 3920 - 0001 *(and)* ISSN - 2583-2344
Published by
Institute of Legal Education
https://iledu.in

3. Methodology

This research employs a qualitative approach combining policy analysis, case study examination, and secondary data analysis. The methodology was designed to capture both the formal regulatory landscape and the lived experiences of marginalized communities navigating these frameworks.

First, a comprehensive analysis of India's current legal frameworks governing online speech was conducted, including the Information Technology Act, 2000 (as amended), the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021, and relevant provisions of the Indian Penal Code. This analysis examined the text, implementation processes, and judicial interpretations of these regulations.

Second, the study analyzed five case studies representing different manifestations of the tension between hate speech regulation and free expression for marginalized communities. These cases were selected to represent diverse identity dimensions, including caste, religion, gender, sexuality, and tribal identity. Each case was examined through available court documents, media coverage, and public statements from involved parties.

Third, the research incorporated secondary data from civil society reports, academic studies, and platform transparency reports to identify patterns in content moderation practices and their impacts on marginalized communities. This included analyses of content removal patterns, account suspensions, and appeals processes.

The research acknowledges limitations including the challenge of accessing comprehensive data on content moderation decisions and the difficulty of capturing the full diversity of marginalized communities' experiences in India. Future research would benefit from primary data collection directly engaging affected communities.

4. Regulatory Landscape and Challenges

4.1 Current Regulatory Framework

India's approach to regulating online speech operates through multiple legal mechanisms. Section 153A of the Indian Penal Code criminalizes promotion of enmity between different groups on grounds of religion, race, place of birth, residence, language, etc.[11] Similarly, Section 295A penalizes deliberate acts intended to outrage religious feelings[12]. In the digital realm, Section 69A of the Information Technology Act empowers the government to direct blocking of online content on grounds including preserving public order[13].

The Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021 represent the most recent significant development in this landscape. These rules require intermediaries to remove content that is "defamatory," "obscene," "invasive of another's privacy," or "racially or ethnically objectionable" within specified timeframes[14]. The rules have been criticized for their broad language and limited procedural safeguards.

4.2 Challenges in Implementation

The implementation of these regulatory frameworks presents several challenges for marginalized communities. First, ambiguity in legal definitions of prohibited speech creates uncertainty about what constitutes legitimate expression. Choudhary (2022) documents instances where content expressing marginalized caste perspectives on historical discrimination was removed for allegedly "promoting hatred," highlighting how contextual factors are often overlooked[15].

Second, procedural aspects of content regulation frequently disadvantage vulnerable users. Redressal mechanisms typically require digital literacy, time, and resources that marginalized communities may lack. A study by the Centre for Internet and Society (2021) found that users from rural areas and lower socioeconomic backgrounds experienced significantly longer resolution times for appeals

against content removal[16]. Third, the increasing use of automated content moderation systems presents particular challenges. Research by Sambasivan et al. (2021) demonstrates that these systems often fail to recognize culturally specific forms of harmful expression while simultaneously flagging legitimate cultural and political expression from marginalized communities[17]. This technological bias compounds existing structural disadvantages.

5. Case Studies and Analysis

5.1 Case Study: Dalit Expression and Anti-Caste Discourse

The experience of Dalit activists and commentators online illustrates the complex dynamics of speech regulation. In multiple documented instances, content describing lived experiences of caste discrimination has been removed from platforms like Facebook and Twitter for allegedly violating community guidelines against "hate speech"[18]. Simultaneously, explicitly casteist content often remains accessible despite reporting.

This asymmetric moderation reflects a failure to contextualize speech within existing power structures. As Ayyub (2021) argues, "When the expression of marginalized communities about their oppression is classified as hate speech, it reinforces the very silencing these communities have historically experienced"[19]. This case demonstrates how seemingly neutral content policies can reinforce existing inequalities.

5.2 Case Study: Religious Minority Voices During Communal Tensions

During periods of communal tension, members of religious minority communities face particular challenges in digital expression. Analysis of content moderation patterns during the 2020 Delhi riots revealed that Muslim voices documenting violence were disproportionately removed compared to misleading content promoting anti-Muslim narratives[20].

This case highlights how temporal context affects speech regulation. During sensitive periods, platforms often implement stricter moderation policies that can inadvertently silence documentation of abuses. As Krishnan (2021) notes, "For vulnerable communities, the ability to document and share experiences of violence is not merely expression but a vital safety mechanism"[21].

5.3 Case Study: LGBTQ+ Content and "Obscenity" Regulations

India's LGBTQ+ communities have leveraged digital platforms for advocacy, community-building, and expression. However, research indicates that LGBTQ+ content is disproportionately flagged under regulations targeting "obscenity" and "public morality"[22]. This pattern reveals how cultural biases embed themselves in regulatory frameworks that appear facially neutral.

The experiences of queer content creators demonstrate how ostensibly protective regulations can perpetuate marginalization. As one content creator noted in interview data collected by Agarwal and Panda (2022), "The same content that would be considered educational if about heterosexual relationships is flagged as inappropriate when it addresses queer experiences"[23].

6. Developing Balanced Approaches

6.1 Centering Marginalized Communities in Policy Development

Effective regulation requires centering marginalized communities in policy development processes. This means meaningful consultation with diverse communities, representation in decision-making bodies, and ongoing evaluation of impacts on vulnerable users. Models like South Africa's process for developing digital content regulation, which mandated participation from historically disadvantaged communities, offer instructive examples[24].

6.2 Contextual Approaches to Content Moderation

Context-sensitive content moderation requires investment in human review systems with

appropriate cultural competency and language capabilities. Platforms operating in India must develop more nuanced guidelines that recognize differences between hate speech and legitimate criticism of power structures or documentation of discrimination.

6.3 Community-Based Moderation Models

Community participation in moderation processes represents a promising approach. Kumar and Shah (2023) document experiments with community-based moderation in regional language online communities that have successfully balanced protection and expression concerns[25]. These models involve affected communities in setting standards and reviewing borderline cases, ensuring that contextual knowledge informs decision-making.

6.4 Procedural Justice in Content Governance

Improving procedural aspects of content governance is essential for protecting marginalized voices. This includes transparent notification systems, accessible appeals processes, and remedies proportionate to harms. Research indicates that marginalized users often experience "remedial justice gaps," where available remedies fail to address the specific harms they experience[26].

7. Conclusion

The tension between regulating hate speech and preserving free expression for marginalized communities in India's digital ecosystem requires nuanced approaches that move beyond simplistic trade-offs. This research demonstrates that current regulatory frameworks often fail to adequately protect vulnerable populations while simultaneously restricting their legitimate expression.

Effective solutions must recognize the differentiated impacts of both hate speech and speech regulation on marginalized communities. This requires regulatory approaches that are context-sensitive, participatory, and attentive to power dynamics. It also demands recognition that for many marginalized communities, digital expression represents not just a right but a vital means of challenging historical exclusion and discrimination.

Future research should explore innovative regulatory models that center affected communities, investigate the specific impacts of content moderation on different marginalized groups, and develop metrics for evaluating regulatory success that include expressive justice for vulnerable communities. By advancing more equitable approaches to digital speech regulation, India has the opportunity to develop a digital ecosystem that truly empowers all voices.

References

Abraham, S., & Hickok, E. (2018). Government Access to Private-Sector Data in India. *International Data Privacy Law*, 8(1), 40-52.

Agarwal, P., & Panda, R. (2022). Digital Erasures: Content Moderation and India's Queer Communities. *Journal of Communication Technology*, 14(2), 178-195.

Ayyub, R. (2021). When Speech Silences Speech: The Paradox of Anti-Discrimination Provisions in Indian Digital Regulation. *South Asian Journal of Human Rights*, 12(3), 215-233.

Bhatia, G. (2021). Offend, Shock, or Disturb: Free Speech under the Indian Constitution. Oxford University Press.

Centre for Internet and Society. (2021). Unheard Appeals: An Analysis of Content Moderation Appeals in India. CIS Research Report.

Chima, R. J. S., & Kaur, J. (2022). Caught in the Net: Content Regulation in India. *Internet Policy Review*, 11(2), 1-24.

Choudhary, S. (2022). The Politics of Silence: Content Takedowns and Anti-Caste Expression in Digital India. *New Media & Society*, 24(6), 1441-1458.

Gupta, T. (2020). Coordinated Campaigns Against Religious Minorities on Indian Social Media. Digital Rights Foundation.

Internet Democracy Project. (2021). Intersectional Harassment Online: A Study of Dalit Women's Experiences on Social Media. Research Report.

Kovacs, A., & Ranganathan, N. (2019). Data Justice: A Study of Platform Content Regulation and Its Impacts on Digital Rights in India. Internet Democracy Project.

Krishna, S. (2022). Platform Governance in India: A Comparative Analysis of Approaches. *Policy Studies Journal*, 50(3), 612-635.

Krishnan, K. (2021). Digital Witnessing and the Limits of Content Moderation During Communal Violence. *Information Technology & People*, 34(6), 1872-1890.

Kumar, S., & Shah, N. (2023). Community Standards: Experiments in Participatory Content Moderation in Indian Language Digital Spaces. *New Media & Society*, 25(4), 809-827.

Ministry of Electronics and Information Technology. (2021). The Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021. Government of India.

Narrain, S. (2016). Hate Speech, Hurt Sentiment, and the (Im)Possibility of Free Speech. *Economic & Political Weekly*, 51(17), 119-126.

Sambasivan, N., Arnesen, E., Hutchinson, B., Doshi, T., & Prabhakaran, V. (2021). Re-imagining Algorithmic Fairness in India and Beyond. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 315-328.

Endnotes

1. Krishna, S. (2022). Platform Governance in India: A Comparative Analysis of Approaches. *Policy Studies Journal*, 50(3), 612-635.
2. Internet Democracy Project. (2021). Intersectional Harassment Online: A Study of Dalit Women's Experiences on Social Media. Research Report.
3. Narrain, S. (2016). Hate Speech, Hurt Sentiment, and the (Im)Possibility of Free Speech. *Economic & Political Weekly*, 51(17), 119-126.
4. Bhatia, G. (2021). Offend, Shock, or Disturb: Free Speech under the Indian Constitution. Oxford University Press.
5. Narrain, S. (2016). Hate Speech, Hurt Sentiment, and the (Im)Possibility of Free Speech. *Economic & Political Weekly*, 51(17), 119-126.
6. Kovacs, A., & Ranganathan, N. (2019). Data Justice: A Study of Platform Content Regulation and Its Impacts on Digital Rights in India. Internet Democracy Project.
7. Internet Democracy Project. (2021). Intersectional Harassment Online: A Study of Dalit Women's Experiences on Social Media. Research Report.
8. Gupta, T. (2020). Coordinated Campaigns Against Religious Minorities on Indian Social Media. Digital Rights Foundation.
9. Abraham, S., & Hickok, E. (2018). Government Access to Private-Sector Data in India. *International Data Privacy Law*, 8(1), 40-52.
10. Chima, R. J. S., & Kaur, J. (2022). Caught in the Net: Content Regulation in India. *Internet Policy Review*, 11(2), 1-24.
11. Indian Penal Code, 1860, Section 153A.
12. Indian Penal Code, 1860, Section 295A.
13. Information Technology Act, 2000, Section 69A.
14. Ministry of Electronics and Information Technology. (2021). The Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021. Government of India.
15. Choudhary, S. (2022). The Politics of Silence: Content Takedowns and Anti-Caste Expression in Digital India. *New Media & Society*, 24(6), 1441-1458.

16. Centre for Internet and Society. (2021). Unheard Appeals: An Analysis of Content Moderation Appeals in India. CIS Research Report.
17. Sambasivan, N., Arnesen, E., Hutchinson, B., Doshi, T., & Prabhakaran, V. (2021). Re-imagining Algorithmic Fairness in India and Beyond. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 315-328.
18. Choudhary, S. (2022). The Politics of Silence: Content Takedowns and Anti-Caste Expression in Digital India. *New Media & Society*, 24(6), 1441-1458.
19. Ayyub, R. (2021). When Speech Silences Speech: The Paradox of Anti-Discrimination Provisions in Indian Digital Regulation. *South Asian Journal of Human Rights*, 12(3), 215-233.
20. Krishnan, K. (2021). Digital Witnessing and the Limits of Content Moderation During Communal Violence. *Information Technology & People*, 34(6), 1872-1890.
21. Krishnan, K. (2021). Digital Witnessing and the Limits of Content Moderation During Communal Violence. *Information Technology & People*, 34(6), 1872-1890.
22. Agarwal, P., & Panda, R. (2022). Digital Erasures: Content Moderation and India's Queer Communities. *Journal of Communication Technology*, 14(2), 178-195.
23. Agarwal, P., & Panda, R. (2022). Digital Erasures: Content Moderation and India's Queer Communities. *Journal of Communication Technology*, 14(2), 178-195.
24. Krishna, S. (2022). Platform Governance in India: A Comparative Analysis of Approaches. *Policy Studies Journal*, 50(3), 612-635.
25. Kumar, S., & Shah, N. (2023). Community Standards: Experiments in Participatory Content Moderation in Indian Language Digital Spaces. *New Media & Society*, 25(4), 809-827.
26. Kovacs, A., & Ranganathan, N. (2019). Data Justice: A Study of Platform Content Regulation and Its Impacts on Digital Rights in India. Internet Democracy Project.