International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Policies on Social Media to Control Religious and Caste Hate Speech: Need Vs Misuse Risks.

Nirmala Tripathi

Associate Professor, Social Science, RKDF, University, Bhopal (M.P.)

ABSTRACT

This paper examines India's social media policies to control religious and caste-based hate speech, weighing their necessity against misuse risks. With over 800 million users by 2025, platforms like WhatsApp and Facebook amplify divisive content, leading to real violence as in the 2020 Delhi riots (53 deaths). Laws such as IT Rules 2021 and Karnataka Hate Speech Bill 2025 enable swift content removal but suffer from vague definitions and selective enforcement, chilling free speech under Article 19(1)(a) per Shreya Singhal (2015). Doctrinal analysis of statutes, cases, and reports reveals enforcement bias and overreach. Reforms—precise definitions for imminent harm, contextual review, and oversight—offer balance to protect harmony without democratic harm.

Keywords: Hate speech, social media regulation, free speech, religious discrimination, caste discrimination, IT Rules 2021, Karnataka Hate Speech Bill 2025, misuse risks, doctrinal analysis

INTRODUCTION

India's social media world has grown huge. By 2025, over 800 million people use it—almost 60% of the country's people. This big change has made hate speech on religion and caste spread very fast. It often turns into real fights and breaks society apart. Apps like WhatsApp, Facebook, and X help bad messages go viral quick. They target poor groups.

Bad Real-Life Cases

The 2020 Delhi riots resulted in 53 fatalities and the displacement of numerous individuals. Investigations by the Delhi Police identified WhatsApp forwards and Facebook posts inciting enmity between Hindu and Muslim communities as key triggers, with courts examining over 1,000 such messages. Concurrently, anti-Dalit vitriol escalated, as evidenced by CSOHate documenting more than 2,500 abusive Facebook posts from 2023 to 2025, which correlated with rural lynchings in states including Uttar Pradesh and Bihar.

Legal Framework for Control

India uses strong laws to fight online hate speech. The Information Technology Act, 2000 (Section 69A) lets authorities block harmful content that threatens public order. The IT (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021, require large social media platforms (over five million users) to appoint grievance officers, monitor content actively, and remove hate speech within 36 hours of complaints, or lose safe harbour protection under Section 79. Indian Penal Code sections—153A (promoting enmity between groups, up to three years in jail) and 295A (outraging religious feelings, similar penalty)—support these efforts. New state laws, like the Karnataka Hate Speech and Hate

Crimes (Prevention) Bill, 2025, add group liability, 2-10 year sentences for repeat offenses, and limits on private message forwards.

Issues and Court Fights

However, these regulations face significant challenges due to vague terminology, such as "hate speech," which risks curtailing legitimate expression. The Supreme Court in Shreya Singhal v. Union of India (2015) struck down Section 66A of the IT Act as overly broad, safeguarding free speech under Article 19(1)(a) and mandating narrow tailoring with due process. In 2024, selective enforcement removed satirical posts by opposition figures while permitting inflammatory content from ruling affiliates, fostering bias, self-censorship among journalists, and erosion of democratic discourse.

Path Forward

Social media regulations play a vital role in protecting India's diverse cultural fabric from the corrosive effects of online hate speech. However, equitable enforcement is essential to prevent authoritarian overreach and misuse of authority. Incorporating precise definitions—such as speech inciting imminent violence—ensures a balanced approach that upholds free expression under Article 19(1) (a) while mitigating tangible harms.

LITERATURE REVIEW

David's 2025 article looks at how AI systems in India carry hidden biases that hurt fairness. Fast AI growth mixes with old problems like caste, religion, gender, and money gaps from history. AI built without input from poor groups keeps these wrongs alive and harms them more. The study uses ideas from bias theory, tech-society links, and women's views on unfair power. It talks to five experts in school, law, science, and rules. Key points: AI data skips out poor groups; no clear ethics rules; spy tools misused; makers and leaders don't know ethics well. Fixing bias after it happens makes things worse. Solutions push fair data, make AI builders take blame, add watch groups, and get people talking on ethics. Focus on fairness lets AI help all without hurting rights.

Sony's 2024 paper explains how digital media growth lets anyone create and share content easily, changing entertainment worldwide, especially in India with its huge variety of viewers. But too much online content brings problems like fake news, copyright theft, and cultural hurt. India faces tough choices in making rules for this new world. The paper looks at current laws and struggles to balance free speech with India's many cultures. Smart rules can stop lies, guard safety and privacy, bring people together, cut fights between groups, and give fair chances to all. It checks three cases: Swami Ramdev's 2019 suit against Facebook, the 2021 Tandav web show fight, and the 2020 TikTok ban over privacy fears. It compares India's rules to those in the US, Europe, Australia, and China. In the end, it calls for better laws and steps to protect users, keep culture safe, and ensure good content.

Ermida's 2023 chapter starts by defining hate speech clearly, as experts often disagree on what it means—some say it's too vague or covers everything. This matters for tech tools like language processing and laws that spot and stop it. Datasets for hate speech depend on good definitions, just like telling it apart from free speech. Using a basic communication model—senders, messages, channels, receivers—the chapter adds new language ideas to make a five-factor test: when all fit together, it's hate speech. It checks the model on real examples from a dataset called NETLANG, covering sexism, racism, and ageism. The chapter also looks at other language tricks in online hate. Still, spotting hate speech fully needs to see how people use it in real talks for certain listeners at certain times.

Binny's 2022 study shows how COVID-19 in India worsened old hate against Muslims and Dalits. Muslims faced strong blame on TV and social media as virus spreaders, especially after a Tablighi Jamaat event in March 2020. Right-wing leaders called them anti-national "super-spreaders," making hidden Islamophobia very public. Dalit hate was sneakier: WhatsApp and Facebook posts praised caste rules and untouchability as "social distancing." This painted Dalit areas as dirty disease spots, costing thousands of low-caste workers their jobs when upper-caste families fired them. The research checks news reports and talks with 100 people in hate-spreading social media groups.

Banaji's 2021 book *Social Media and Hate* explores how misinformation and hate speech on platforms like Facebook, TikTok, Instagram, WhatsApp, and ShareChat fuel violence and discrimination. Through expert interviews, focus groups, theory, data studies, and cases from India, Brazil, Myanmar, and the UK, it shows links between users' politics, values, and targeted attacks on groups like Muslims, Dalits, feminists, LGBTQIA, Rohingya, and immigrants. The book stresses similar tech and idea patterns across contexts, urging strong political fixes for scholars in media, politics, psychology, and sociology.

Sinpeng's 2021 study, funded by Facebook, checks hate speech laws in India, Myanmar, Indonesia, the Philippines, and Australia. It compares expert definitions of hate speech with Facebook's own rules. The research looks at how Facebook handles content by reading company papers, talking to staff, and tracking responses. It studies hate posts on public Facebook pages of LGBTQ+ groups that slipped past filters. Suggestions for Facebook include talking more with affected groups, making a clear way for crisis reports from trusted partners, holding yearly meetings with local experts, and training page admins as key watchers to fight hate better.

Wilson's 2021 article says social media firms banning hate speech settles old US fights on limits—now more rules win. But old talks give key lessons for smart online rules. First, check if hate truly causes violence with real data from platforms. This helps make fair policies that hit real harm without too-wide bans. Second, company rules ignore context, like local power gaps and politics. Hate's harm depends on who says it and where. Without that, rules over-block. The piece links online hate to strong nationalism and attacks on minorities. It suggests adding context to moderation for better fixes.

Mirchandani's 2019 paper looks at how social media changes news, talks, and info sharing in India. It helps free speech for all but also lets lies and hate grow under "free speech." This supports big-group violence as leaders push "us vs. them" politics. The study asks if anti-terror groups can fight Hindu majoritarian hate too. A 2018 report found religion and food/dress hate rose from 19% to 30% on Indian social media pages. Most targeted Muslims (180 million people), pushing harm over marriages, rights, cows, and beef. Social media speeds up "offended" feelings, like Hindu far-right in India matching others worldwide, turning online hate into real fights.

Balkin's 2018 essay says we live in the Algorithmic Society. Algorithms, AI, and Big Data let digital firms like social media control speech between governments and users. This causes problems: firms trick users with data and hide behind free speech to skip rules. Platforms now rule speech like old states, but laws don't protect well. Fixes need honest actions from info firms and no harm-dumping. Now, governments, users, and groups pull platforms in a "pluralist" fight over rules, as in fake news cases. Users demand fair play, clear steps, and steady norms from these speech rulers.

Chetty's 2018 paper explains that internet advances and social networks offer great benefits but also boost hate speech and terrorism worldwide. Hate speech spreads mean ideas using stereotypes about gender, religion, race, or disability, while terrorism targets lives and safety. Often discussed apart, hate speech acts as terrorism's spark after events, thriving on youth-popular platforms. The review covers

hate types and online terror tactics, urging governments, internet providers, and networks to unite on smart policies for effective control.

Pohjonen's 2017 article studies India and Ethiopia to explain "extreme speech" on internet media. It looks at mean online fights. Most call it "hate speech," but this view has problems. First, extreme speech needs real context—like user habits and local talk history. Second, online meanness is not black-and-white; it's hard to split hate from okay speech. The paper uses "comparative practice" to check this. It fights simple ideas of internet danger that push speech bans.

Roy's 2017 study uses two ideas—Spiral of Silence and Securitization—to check how the US and India governments remove online content from sites like Google, YouTube, Facebook, and Twitter. This acts like internet control for safety. The research mixes interviews with seven experts on removal rules from 2010-2015 and surveys of 587 people in both countries about self-censorship on Facebook and Twitter, especially on security and government criticism. Findings show US removals often cite "defamation" wrongly, while India uses "religious offense" and "defamation." Both delete safe free speech. This government action makes people scared to post on hot topics, boosting self-silence online. Using laws against critics hurts open talk in democracy.

Research Methodology

This study uses a qualitative doctrinal approach based solely on secondary data. It examines social media policies to control religious and caste-based hate speech in India, focusing on the balance between need and misuse risks under laws like IT Rules 2021 and Karnataka Hate Speech Bill 2025.

Data Sources

Data come from these main types:

- Laws and rules: IT Act 2000 (Section 69A), IT Rules 2021, IPC Sections 153A and 295A, Karnataka Bill 2025.
- Court cases: Shreya Singhal v. Union of India (2015) and 2024 enforcement examples.
- Reports: CSOHate (2023-2025), ORF (2018), Law Commission Report 267 (2017).
- Academic works: Summaries from David (2025), Sony (2024), Banaji (2021), Wilson (2021), and others in the literature review.

Data Collection and Analysis

Data cover 2015-2025. Sources include academic databases like Google Scholar and JSTOR, plus official sites such as India Code and Supreme Court records. Purposive sampling selects relevant items on hate speech policies and cases like Delhi riots (2020). Thematic analysis finds patterns in enforcement bias, vague terms, and free speech limits. Triangulation checks data across types for trust. Theme coding, like "selective use" or "context gaps," follows standard methods. This desk study fits policy analysis well. It stays objective without new data collection.

Findings

Analysis of secondary data shows three main patterns in India's social media policies for religious and caste hate speech. First, enforcement is uneven. Platforms often remove opposition posts but ignore ruling party content. CSOHate data from 2023-2025 found over 2,500 anti-Dalit Facebook posts linked to violence in Uttar Pradesh and Bihar, yet few led to removals. Second, vague terms like "hate speech" cause problems. The Supreme Court in Shreya Singhal v. Union of India (2015) struck down Section 66A of the IT Act for being too broad, protecting free speech under Article 19(1)(a). Third, online hate leads to real harm. The 2020 Delhi riots killed 53 people after over 1,000 WhatsApp and Facebook messages spread Hindu-Muslim hate.

Table 1: Main Enforcement Patterns

Pattern	Examples	Impact
Selective removal	2024: Opposition satire deleted; ruling party hate stayed	Builds bias and chills journalism
Vague definitions	IT Rules 2021 lack clear "hate speech" line	Stops normal talk; courts overturn broad laws
Online to offline harm	Delhi riots (2020): 1,000+ messages; 53 deaths	Shows need for fast action

Discussion

These findings highlight a balance problem. Policies meet real needs in a country with 800 million social media users where hate spreads fast. Events like Delhi riots prove quick removal under IT Rules 2021 and Karnataka Bill 2025 can save lives. Yet misuse risks grow, as Roy (2017) found self-censorship after government takedowns. Wilson (2021) adds that rules must consider context—who speaks, where, and power gaps—to avoid over-block.

Literature supports reform. Banaji (2021) links user politics to targeted hate on Dalits and Muslims. Pohjonen (2017) calls for local speech history in definitions. A fix: limit rules to speech causing instant violence, per Law Commission Report 267 (2017). Add oversight boards for fair checks.

Table 2: Need vs. Misuse Risks

Aspect	Need (Why Policies Help)	Misuse Risks (Problems)	Suggested Fix
Speed	Removes 1,000+ riot messages in 36 hours	Deletes satire as "hate"	Clear rules for imminent harm only
Reach	Covers WhatsApp forwards and big platforms	Ignores context like power imbalance	Train officers on Article 19 limits
Scale	Handles 800M users and caste/religion fights	Self-censorship scares writers	Independent review boards

This approach protects India's diverse groups while keeping free speech safe.

Conclusion

India's social media policies on religious and caste hate speech strike a delicate balance between urgent need and misuse risks. With over 800 million users fueling rapid spread of divisive content, laws like IT Rules 2021 and Karnataka Hate Speech Bill 2025 enable swift action against harms, as seen in the 2020 Delhi riots where online incitement led to 53 deaths. Yet, vague definitions and selective enforcement—evident in 2024 removals of opposition satire while ruling party content persists—threaten Article 19(1)(a) free speech rights and foster self-censorship.

This doctrinal analysis of statutes, cases, and reports underscores the necessity of precise reforms: limit "hate speech" to content inciting imminent violence, mandate contextual review, and establish independent oversight. Such measures, aligned with Shreya Singhal (2015) and Law Commission Report 267 (2017), safeguard communal harmony without democratic erosion. Future research should explore primary enforcement data to test these proposals empirically.

References

1. **Balkin, J.M.** (2018) *Free speech is a triangle*, Columbia Law Review, pp. 2011–2056.
2. **Banaji, S.** (2021) *Social media and hate*. Routledge.
3. **Binny, K.** (2022) *Hate speech against Muslims and Dalits during COVID-19: A study of Indian social media*.
4. **Chetty, N.** (2018) *Hate speech and terrorism on online social networks: A review*.
5. **CSO Hate** (2025a) *Online abuse against Dalits: 2023-2025 report*.
6. **CSOHate** (2025b) *Selective enforcement on social media platforms*.
7. **David, ?** (2025) *Algorithmic bias and discrimination in India*.
8. **Drishti IAS** (2025) *Delhi riots 2020: Social media role*.
9. **Ermida, I.** (2023) *Hate speech identification: A communication model*.
10. **Kaur, ?** (2025) *Hate speech trends in India 2025.*
11. **Law Commission of India** (2017) *Report 267: Hate speech definitions*. Government of India.
12. **Mirchandani, V.** (2019) *Digital hatred, real violence: Majoritarian radicalisation and social media in India*, ORF Occasional Paper #167. Observer Research Foundation.
13. **Next IAS** (2025) *Karnataka Hate Speech Bill 2025 analysis*.
14. **Observer Research Foundation (ORF)** (2018) *Hate speech mapping on Indian social media*. New Delhi: ORF.
15. **Pohjonen, N.** (2017) *Extreme speech as a label for online vitriol*, New Media & Society.
16. **Roy,** (2017) *Spiral of silence and securitization: Government content removal in US and India*.
17. **Shreya Singhal v. Union of India** (2015) Writ Petition (Criminal) No. 167 of 2012, Supreme Court of India.
18. **Sinpeng, A.** (2021) *Hate speech law mapping: India, Myanmar, Indonesia, Philippines, Australia*. Facebook Content Policy Research Awards.
19. **Sony, ?** (2024) *Regulatory challenges in digital media entertainment: India focus* .
20. **Vajiram & Ravi** (2025) *State innovations in hate speech laws*.
21. **Wilson, B.** (2021) *Hate speech regulation in the platform age*.