# Profitable Provocations:

# A Twitter-based Study of Abuse and Misogynistic Trolling Directed at Indian Women in Public-political Life

## A summary of emerging findings and policy recommendations

IT for Change

April 2022

Undertaken by IT for Change under 'Recognise, Resist, Remedy' – a project supported by IDRC, Canada.

# 1.Introduction

Over the past decade, there has been a growing awareness of the ways in which social media platforms have engineered a fundamental shift in the traditionally understood notion of the public sphere. Mounting evidence has shown that the harmful societal effects of the platformized public sphere have been disproportionately borne by women and marginalized groups, as viral hate and misogyny is allowed to spread almost completely unhindered.

Online violence against women has become normalized and routinized to such an extent that it has almost been rendered invisible and imperceptible. Outside of specialized technical circles, instances of violence are seen as aberrations or isolated disturbances, and the magnitude of the problem is not fully appreciated. Online violence against women takes a range of forms and expressions, ranging from threats of violence, rape, and murder to seemingly milder forms of trolling by thousands of trigger-happy online trolls.

Discussions about hate speech and its regulation often get mired in irresolvable contradictions and bottlenecked in legal questions such as where to draw the line in the sand in regulating user speech. Set apart from the lofty rhetoric of free speech doctrine, we have adopted a more grounded and ordinary starting point in our study of online violence – its materiality *in situ*. This approach has helped us bring into sharper relief the immense volume of online violence of a *seemingly* milder variety. This sort of violence derives its toxicity and potency not only from its substantive content, but from its volume, frequency, and contextual mode of delivery.

This document presents a summary of the findings and recommendations from our research study on hateful and abusive speech on Twitter directed at 20 Indian women in public-political life.

# 2.Methodology

We used the method of purposive sampling to select women with varied political affiliations -- those formally involved in party politics (both women who are currently holding office as well as those not in office) as well as political commentators engaged in public life such as journalists, and voices of civil society. A high engagement rate on the platform was deemed an important criterion in sample selection as it would provide a densely populated dataset that could be used to draw substantiable and robust inferences. The endeavor was to create a representative sample across the axes of caste, age and geography, and also along the ideological spectrum.

We used the API provided by the Twitter developer platform to collect data from their public profiles, and inductively developed a set of annotation guidelines to annotate the mentions, and to classify the problematic mentions into 22 mutually exclusive codes which defined different forms of neutral,

abusive or problematic speech. Some examples of problematic speech codes include generic abuse, delegitimizing by othering, religious hate, casteist hate, derailing, and overfamiliarity. The 30,000-odd mentions that were addressed to the women in our sample, over the one-week period of data collection, were annotated and categorized under these categories of abusive or problematic speech. By the end of the annotation process, we were left with just under 2,000 abusive, hateful or problematic mentions.

# 3.Findings

The third chapter of the report contains a detailed examination of the various patterns of violence that we were able to discern through an in-depth analysis of the problematic mentions. The text of the final report is interspersed with “snapshots” or verbatim reproductions of abusive tweets (with the handles and other personally identifiable information redacted), as we believe that an unsanitized presentation is necessary to more effectively convey not only the kinds of violent speech acts we reference, but also the gravity of the issue.

The broadest finding from our research is that all the women in our sample, regardless of whether they belong to the opposition or ruling parties, whether they are perceived to be dissenters or sympathetic to the current dispensation, received some amount of abuse on the platform. None of the women in our sample were entirely spared. This disabused us of the notion that all trolls are right-wing users. However, those who were perceived to be ideologically left-leaning, dissenters, and women from opposition parties received a disproportionate amount of abusive and hateful messages. Muslim women and women political commentators who did not have formal party protection, especially, were at the receiving end of an inordinate amount of abuse.

Another, perhaps counter-intuitive, but important finding has to do with the most common kinds of gendered abusive speech on the platform. While we did certainly come across many abusive messages that were threats of violence, wished death on women, and other messages of a similarly grave nature, far more common were forms of trolling of a supposedly milder variety- in the nature of tongue-in-cheek jokes and remarks. We found this ‘fun’ culture of vitriol and abuse to be rampant on the platform, through the sharing of misogynistic memes and word-play.

We also found that the abusive speech received by women in public-political life rarely had anything to do with their work or their politics. It invariably took the form of gendered attacks on their bodies or character. Rather than responding to or engaging with (even with anger or abuse) political positions, trolls usually resorted to either questioning women’s credentials, or trivializing their role in politics. This was done in a variety of ways, usually by insinuating that women were just there to make up the numbers and that the real authority lay with the men in their parties/family. A very common tactic was

derailing and whataboutery, where an entirely irrelevant incident or event in the life of the woman would be brought up to discredit her. There were also numerous irrelevant comments made about the women's appearance and how they were "all beauty and no brains".

Another highly generalizable pattern we found had to do with how trolls targeted particular users and posts. It was not the case that every post by a woman was treated equally by trolls. Perhaps partly as a way of establishing some kind of fraternal solidarity, and partly as a way to exploit the system of algorithmic amplification on Twitter, we found the prevalence of a kind of herd aggression where users banded together to reply to only certain posts. In addition, we also found that trolls tended to tag and abuse certain women together, based on their supposed membership in imagined groups such as the "*tukde tukde gang"*[1], or women in politics who have a background in the entertainment industry.

With regard to some of the more specific sub-continental misogynistic tropes that were commonly invoked, we found there to be a clearly discernable preoccupation with the much written about *shame/honor* trope. "Shame on you" or "hang your head in shame" and its equivalent in Indian vernacular languages was a most common refrain, often used to convey that the woman had not only damaged her own reputation (for whatever perceived indiscretion) but had brought shame to her husband, father, party, community or nation. There were also numerous mentions that sexually objectified women and viewed them solely as the objects of male desire or disgust. There were many variations on this theme, but what all of these had in common was an idealized view of womanhood that does not exist outside of the patriarchal imagination within which they are either placed on a pedestal and worshipped, or only worthy of being treated with disdain.

In addition, we also found that the nature of abusive messages received by women in our sample varied significantly depending on their membership in marginalized communities. Our findings seem to suggest that casteist stereotyping and insinuations of casteism were more common than the use of direct casteist slurs. One common stereotype was the familiar theme of calling into question the merit of the lower-caste women, through the use of words such as *aukaat* (loosely translates to status/ability) and *ghotalebaaz* (scamster). Allegations of corruption were also disproportionately leveled against lower-caste politicians, casting aspersions on their professional integrity.

It is difficult to overstate the sheer volume of abusive and hateful mentions directed at the Muslim women in our sample. Broadly speaking, our findings show that most of these mentions fell into two buckets – first, hypernationalistic and othering speech that attacked them on the basis of a narrow definition of nationality or by asserting an exclusive claim over national identity. The second broad

[1]*Tukde Tukde Gang* is a pejorative political catchphrase used in India by the ruling Bharatiya Janata Party (BJP) and its sympathisers accusing their critics for allegedly supporting sedition and secessionism. See: https://en.wikipedia.org/wiki/Tukde_Tukde_Gang

category of mentions expressed stereotypical, hateful, and clearly propaganda-informed generalizations about Islam. We also found the prevalence of inauthentic user behavior in the form of numerous sets of cookie-cutter tweets that were meant to instill fear in the reader, and a sense that the Muslim community poses a threat to the integrity of the nation. “Love jihad” was among the most tweeted about subjects in the period that we collected data, which happened to coincide with the introduction of the anti-conversion bills in some states of the country.

Since the period of data collection was only one week, there is, naturally, a skewed representation of the gamut of political issues. Some of the most intensely debated issues at the time, towards the end of November 2020, were the ongoing farmer protests, the untimely death of Bollywood actor Sushant Singh Rajput, the West Bengal Assembly elections, and the enactment of anti-conversion laws in various states. Given the constant churn of political content and the ways in which virality dictates engagement on Twitter, these divisive topical issues dwarfed any other political issues that may have arisen at the time.

# 4.Analysis

Based on our findings and the wide gamut of abusive behaviors directed at women in our sample, a central claim that we forward in this report is that in order to grasp the magnitude of the problem of gender-based violence on social media, there is a need to destabilize, unsettle, and problematize the concept of online gendered violence. The first part of this chapter attempts to understand online misogyny and the ways in which the affordances of social media platforms have engendered new speech practices and distinct forms of sociality through two convergent approaches. First, to what extent does the online public sphere *reproduce* the unequal gender relations of the offline world? And second, how do the affordances of social media platforms produce new vulnerabilities and unprecedented ways to target women online?

Responding to the former question, we refer to Uma Chakravarti and Veena Das’s writings on gendered violence,[2] and then to William Mazzarella and Raminder Kaur’s writings on censorship,[3] and arrive at the judgement that, despite the radically new communicational context of social media, there remains, in some important respects, a striking resemblance between the analog and the digital, in the practices of social and cultural regulation that seek to police women’s behavior in the public sphere.

[2]Chakravarti, U. (1993, April 3). *Conceptualising Brahmanical Patriarchy in Early India: Gender, Caste, Class and State*. Economic and Political Weekly, Vol. 28, Issue No. 14. Das, V. (2008, June). *Violence, Gender and Subjectivity*. Annual Review of Anthropology.
[3]Kaur, R., & Mazzarella, W. (2009). *Censorship in South Asia: Cultural Regulation from Sedition to Seduction*, Chapter 1. Between Sedition and Seduction: Thinking Censorship in South Asia.

By studying abusive speech directed at women in public-political life, we are uniquely positioned to be able to see the linkages between the culturally specific forms of gendered regulation (which we elaborate under the conceptual umbrella of Brahmanical patriarchy), and ideological constructions of the nation. Much of the gendered violence that we observed was couched in terms of a muscular and exclusionary Hindu nationalism, and the persistent thread of male control over women's sexual autonomy cut across all experiences of misogyny, without exception. Our findings, therefore, seem to corroborate much of the historical feminist scholarship in India regarding the intimate links between the construction of the nation as an imagined community, and the appropriate role of gendered subjects therein.

Mazzarella and Kaur's work on censorship also helped us locate in our own findings, the simultaneous operation of seemingly incommensurate forms of cultural regulation. These practices, of publicity and censorship, incitement and containment, which tended to constellate largely around restraining the sexual and political autonomy of women, were clearly observable in the way that social media controversies routinize a pattern of incitement and reward bad-faith actors.

To the second question, we begin by attempting to sketch out a profile of the political online troll, and then move outward into the surrounding social media milieu he comes to compose. In this context, we discuss visions of online participatory democracy where ordinary citizens can make their voices heard through collective online engagement and meaningfully contribute to the field of politics. We refer to C. Thi Nguyen's writings on the gamification of public discourse[4] to highlight the ways in which platform design plays a determinative role in shaping online interactions, by incentivizing shallow and provocative engagement. We then refer to Sahana Udupa's work on the links between the aspirations of political participation and online abuse. Her conception of *gaali* (Hindi for abuse) to capture the "interlocking practices of insult, comedy, shame and abuse that unfold in a blurred arena of online speech" was particularly useful to think with in the context of our findings.[5]

In the second part of chapter 4, we turn to the role of the platform in larger socio-political formations. We discuss the tactical collusions between authoritarian states, and the legitimacy afforded to acts of public majoritarian violence. We also consider the ways in which acts of gendered violence are connected with the promise of inclusion into an empowered majority, and with immunities from prosecution.[6] This section situates our findings within the growing body of scholarship on forms of

[4] Nguyen, C.T. (2021). *How Twitter gamifies communication.* Applied Epistemology. Oxford University Press. pp. 410-436.
[5] Udupa, S. (2017). *Gaali cultures: The politics of abusive exchange on social media.* New media & society 2018, Vol. 20(4) 1506-1522.
[6] Hansen, T.B. (2021). *The Law of Force: The Violent Heart of Indian Politics.* Aleph Book Company.

mediated populism in postcolonial democracies and the complex ways in which networked citizens negotiate dominant political imaginaries.[7]

We then conclude the chapter by discussing the inadequacies of the industrial moderation model[8] used by large social media platforms, specifically in the context of regulating online gender-based violence. Citing data from platform compliance reports, we highlight that the policy area where Twitter receives the most user complaints and the policy area where proactive monitoring tools are least effective in detecting harmful content are the same- abuse and harassment. Based on this, we point to the need to devote more resources toward contextually sensitive approaches to content moderation, both at the level of detection- in training local human moderators and developing machine learning tools; as well as improved reporting mechanisms and responsiveness to user grievances. We also argue that user complaints must be utilized as key resources for platforms to fill in the gaps where proactive monitoring tools fall short, such as in the detection of abuse and harassment on the platform.

## 5.Legal-institutional Responses

We begin this chapter by discussing the arguments both for and against the recognition of online gender-based violence in law, and situate this debate within the longer history of feminist debates regarding the role of the law in the regulation of sexual and misogynistic speech.[9] We argue, relying on insights from feminist speech act theory, that the harms of online gender-based violence should not be understood only in terms of its locutionary/semantic content, but rather, in the larger structural context of legitimating the discriminatory treatment of women in the online public sphere, and therefore, as harms that are relevant to law. We also highlight the need to view social media as both a site of danger as well as pleasure and fulfillment for women. In so doing, we discuss the dangers and unintended consequences of laws that seek to proscribe certain forms of speech, which often contain overbroad definitions of what constitutes an offence, and are therefore are liable to misuse and tend to reinforce the marginality of vulnerable groups by reinscribing gender and sexual hierarchies.

Based on the above, we suggest that the way to think about the regulation of online misogyny on social media is a move away from criminal, carceral, and retributive notions of justice; and towards a model of accountability that foregrounds the effective delivery of justice in ways that are responsive to

[7] Govil, N., & Baishya, A.K. (2018). *The Bully in the Pulpit: Autocracy, Digital Social Media, and Right-wing Populist Technoculture.* Communication Culture & Critique 11, 67–84. Srivastava,S. (2015). *Modi Masculinity: Media, Manhood, and "Traditions" in a Time of Consumerism. Television and New Media. Vol 16, Issue 4, 2015.* Pal, J, et al. (2019). *A friendly neighborhood Hindu: Tempering populist rhetoric for the online brand of Narendra Modi.* Conference for Democracy and Open Government.

[8] Caplan, R. (2018). *Content or Context Moderation? Artisanal, Community-reliant and Industrial Approaches.* Data & Society. Gillespie, T. (2018). *Custodians of the Internet: platforms, content moderation, and the hidden decisions that shape social media.* Yale University Press.

[9] Cossman, B. (2021). *The New Sex Wars: Sexual Harm in the #MeToo Era.* New York University Press.

the needs of the victim. We then turn to the question of precisely what kind of legal intervention would be best suited to address the problem of online misogyny and argue that there is a need to move past the individualistic frame of the victim-perpetrator binary,[10] and towards a model of accountability, which holds platforms responsible for the hostile and abusive environments that they foster and profit from. A central component of this argument lies in dismantling the notion of the passive intermediary, or the metaphor of the platform as a "dumb conduit". This involves emphasizing the extent to which platforms *already* mediate content, and thereby regulate speech, not by directly muzzling or stifling speech, but by steering and manipulating the attention of users at a mass scale through algorithmic means, by organizing, ranking, recommending, hiding, and curating content. As Zeynep Tufecki argues, it is less a question of regulating user-generated speech than it is of regulating attention in the public sphere.[11]

Based on these two considerations, first, that online misogyny does not present in easily recognizable ways and would therefore be near impossible to regulate based on its locutionary content alone; and second, in appreciating that *principal* issue which makes online misogyny particularly toxic is not only the content of the messages but their frequency, velocity, and viral spread enabled by the affordances of social media platforms, we make the following recommendations for legal-institutional responses targeted at platforms:

- There is a need for a comprehensive legislation to hold social media platforms liable for online harms (with differential compliance obligations depending on the size of the platform) which imposes on platforms a statutory duty of care to its users, and provides for oversight of platform governance by an independent regulatory authority.
- A central problem in content governance practices across platforms is the lack of consistency with regard to either identification or enforcement action in relation to content flagged under specific policy areas. There is a need to establish a degree of normative coherence to ensure that platform terms of service are enforced predictably, fairly and without arbitrariness. Content governance must be bound by a set of guiding principles in relation to specific policy areas to advance the principle of international comity and harmonize adjudication practices across platforms. Platforms must also make use of consistent terminologies and make efforts to reach a convergence on their terms of service/community standards.
- Especially relevant in the context of tackling online misogyny, a gendered perspective on the right to freedom of expression must include a framework for promoting positive freedoms

[10] Raghavan, A. (2021). *The Internet-Enabled Assault on Women's Democratic Rights and Freedoms*. IT for Change.
[11] Tufecki, Z. (2020). *The Problem With (All) The Tech Hearings.* Insight. https://www.theinsight.org/p/the-problem-with-all-the-tech-hearings

(freedom for) such as the right to public participation, as well as negative freedoms (freedom from). Under the prevailing legal position, negative freedoms largely animate moderation policies and platform design.

- Any measures to address online gender-based violence must meet the standards of legality, necessity, and proportionality. Depending on the nature of the offending content, the regulatory authority shall stipulate different timeframes to take action on the content. The regulatory authority should not be empowered to initiate criminal proceedings, except in egregious cases of harm.
- Comprehensive and regular transparency reporting is essential to more accurately diagnose and address the problem of online gender-based violence. Over and above the bare minimums of the Santa Clara Principles, disclosures must improve transparency in algorithm design and implementation. Transparency reporting formats must be standardized across platforms and provide for disaggregated data according to specific policy areas such as online gender-based violence.
- Platforms must put in place responsive grievance redressal mechanisms for users to submit their complaints. Robust appeal mechanisms are also crucial to offset the significant risk of wrongful decisions by platforms. Both grievance redressal and appellate mechanisms should be easily and quickly accessible to users, and disposed of expediently by platforms.
- Regulatory oversight over moderation practices of platforms must not be limited only to decisions to takedown or reinstate content, but must also extend to design choices and algorithmic processes of amplification of content. For instance, the regulatory authority would be tasked with investigating the practice of shadowbanning by platforms, the use of dark patterns, or ensuring that the trending page on a platform does not cause any hateful content to spread virally. To ensure that AI identification/proactive monitoring meets the standard of normative coherence, tools used must not only analyze content substantively (by looking at keywords or images) but also according to their level of dissemination and virality.