Profitable Provocations

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

JULY 2022

IT for Change

IDRC · CRDI
Canada

AUTHORS

Anita Gurumurthy
Amshuman Dasarathy

This paper is part of IT for Change's project, Recognize, Resist, Remedy: Combating Sexist Hate Speech Online, which aims to address gender-based hate speech in the online public sphere, with support from the International Development Research Centre (IDRC), Canada.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the work of Bhavna Jha in conceptualising and laying the groundwork for the research study; and Nandini Chami for her immensely valuable and significant support at each step of the research process from the data collection phase to the analysis. The study would not have been possible without the meticulous data collection, coding, and pattern analysis undertaken by Faiz Naeem, Harish Nalawade, and Taijshee Mishra.

Copy Editing: Sohel Sarkar Proofing: Intifada P. Basheer & Sadaf Wani
Design & Visuals Support: Sreemoyee Mukherjee
Infographics & Report Design: Indu Manohar

IT for Change

IDRC · CRDI
Canadä

01

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 held notions of the public sphere. With the expansion of internet use the world over, and the continued growth of social media user bases, we have seen the emergence and sustained rise of destructive phenomena, such as coordinated disinformation campaigns and pervasive harassment and abuse on social media platforms.

With their massive user bases across the globe, social media platforms today constitute a vital site for public communication and discourse. Yet, and as is widely recognised today, the profit-oriented business tactics of social media corporations, such as microtargeting and content optimisation, construct an online public sphere that is captive to the all-encompassing logic of the market. Each minute interaction, click, and performative act on social media platforms carries a monetary consequence, which propels a toxic environment wherein users are incentivised to behave in ways that are detrimental to the quality and health of public discourse.

Mounting evidence has shown that the harmful societal effects of the platformised public sphere have been disproportionately borne by women and marginalised groups, as viral hate and misogyny are allowed to spread almost completely unhindered. This is of serious consequence for women's right to public participation and freedom of expression. Platforms rely on different interlocking metaphoric frames as a way to deflect attention away from their own complicity in fostering environments where abuse and misogyny runs rampant.[1] The metaphor of the marketplace of ideas where everyone is benefited by free and unregulated exchange may be an ideal, but it does not exist in practice. Power dictates the operation of the marketplace of ideas. This power is contingent on a range of markers such as religion, caste, class, gender, and sexual orientation. Similarly, the metaphor of the intermediary as a 'dumb conduit,' that only performs a passive hosting or relaying function, renders invisible the extent to which platforms already mediate content algorithmically.

This report presents the findings of a study on the hateful and abusive speech directed at 20 Indian women engaged in Indian public and political life on Twitter.

Discussions around hate speech and its regulation often get bottlenecked in irresolvable contradictions and legal questions, such as where to draw the line in the sand in regulating user speech. Set apart from the lofty rhetoric of the free speech doctrine, our study of online violence adopts a more grounded starting point that foregrounds the everyday acts of online trolling. In so doing, it seeks to lay emphasis on the contextual materiality of online violence, as it routinely occurs in the online public sphere.[2] This approach enables us to bring into sharper relief the immense volume of violence seemingly of a milder variety, often referred to as trolling. Though widely regarded as less serious forms of hostile or violent conduct, trolling derives its toxicity and potency precisely from its sheer volume and frequency, rather than only its substantive content.

Our research points to innumerable such examples of viral trolling, which have pernicious effects and whose harms are not always readily apparent. This kind of speech that is characteristic of online sociality presents unprecedented regulatory challenges. Our central contention in this regard is that it would be erroneous to dismiss these as less serious forms of violence, and that it is necessary to evolve a more sophisticated regulatory approach that engages with the complex issues of virality and amplification of content through online networks.

Online violence against women has become normalised and routinised to such an extent so as to render it almost invisible and imperceptible. It takes a variety of forms and expressions ranging from threats of violence, rape, and murder, to abuse often deemed normal. Yet, outside of specialised technical circles, such instances of online violence are seen as aberrations or isolated disturbances, and the magnitude of the problem is not fully appreciated.

One of our primary motivations in this report is to learn about the patterns of violence that emerge through a systematic analysis of abusive mentions. The study seeks to not only investigate the scale and incidence of online gender-based violence on Twitter, but also to gain insights about the nature of the violence and abusive speech directed at women in public-political life.

A central claim that we put forward is that in order to grasp the magnitude of the problem of online gender-based violence, there is a need to destabilise the notion of violence as an interruption, an aberration or a deviation from the norm. Instead, as Veena Das writes,

> *"The centrality of gender in the understanding of violence will show the deep connections between the spectacular and the everyday [...] There is an increasing public perception that safe havens no longer exist and that peace-time violence is as debilitating as that of war."*[3]

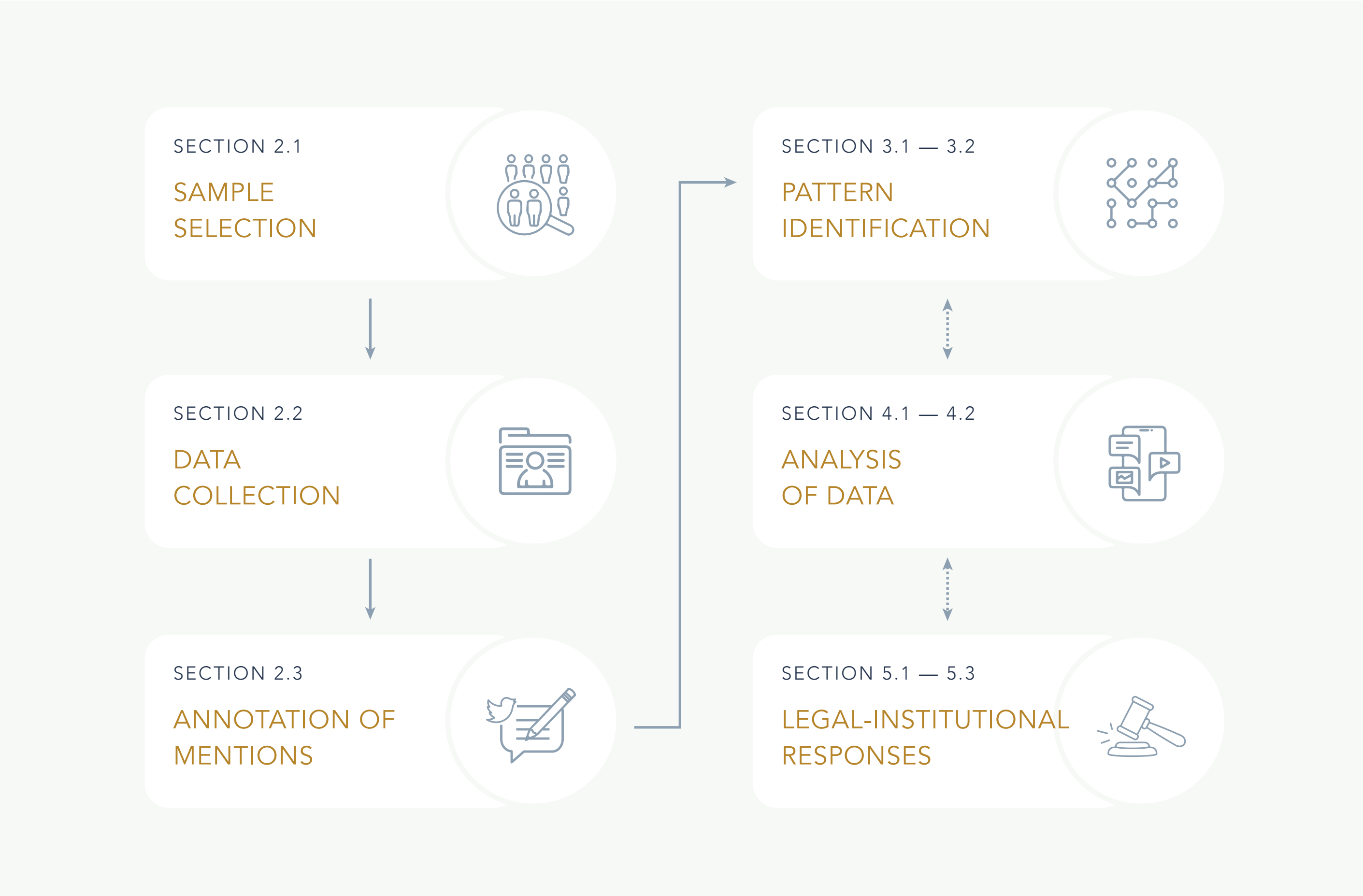
Drawing upon this understanding of violence as an ongoing and continuous process, rather than a state of exception, can help us to recognise that there is a lot at stake in the very act of naming or marking out the boundaries of violence—whether in the traditional feminist understanding of separation of the private and public spheres, or indeed, transposed to the crisscrossing geographies of the online and offline spheres. By drawing attention to the wide range of online behaviours that ought to be recognised as gender-based violence, we argue that the principal issue is not the content of hateful speech, but rather the platformised structures which enable such speech to thrive.

The architecture of social media platforms tacitly encourages and rewards behaviour that is abusive and toxic. Therefore, there is a need to move beyond the victim-perpetrator binary in legal-institutional responses to online violence, and target platforms rather than individual offenders. Rather than seeking to police speech, regulatory intervention must strike at the heart of the epidemic of online violence against women, and contend with the issues of virality, amplification, and coercive deplatforming.

02

Methodology

In order to undertake a systematic analysis of hateful, abusive, and problematic speech on Twitter directed at Indian women in public-political life, we outline the steps of our methodology below.



SECTION 2.1

SAMPLE SELECTION

SECTION 2.2

DATA COLLECTION

SECTION 2.3

ANNOTATION OF MENTIONS

SECTION 3.1 — 3.2

PATTERN IDENTIFICATION

SECTION 4.1 — 4.2

ANALYSIS OF DATA

SECTION 5.1 — 5.3

LEGAL-INSTITUTIONAL RESPONSES

Infographic 1: Research process for this project. ***Click within the graphic to go to the relevant section.***

2.1 Sample selection

Identifying research subjects:

The first step in our research process was to identify a set of women who fit the description of those who are active in public-political life and have a high engagement rate (defined below) on Twitter. A high engagement rate on the platform was deemed an important criterion as it would provide a densely populated dataset that could be used to draw substantiable and robust inferences, and not merely propose speculative hypotheses. Thus, many prominent figures who fit the description of women with an active public-political life have not been included in this sample as they do not have a high engagement rate on the platform.

Subject's engagement rate:

A subject's engagement rate was determined based on two values: the number of followers, and the total number of tweets by the individual. These numbers were normalised on a scale of 0 to 1, where 0 denotes the highest engagement. The engagement rate was arrived at by calculating the average of these two normalised values. Taken together, these parameters helped strike an appropriate balance between inactive accounts with a large following, and accounts with many posts but a small following.

Purposive sampling of subjects:

From the set of individuals arrived at by applying the engagement rate filter, as explained above, we used purposive sampling to select women with varied political affiliations—individuals formally involved in party politics, as well as political commentators engaged in public life such as journalists and activists. Further, we used caste identity, age, and geography as secondary sampling criteria. We also attempted to select women located across the ideological spectrum. The purposive sampling method adopted for this research, certainly does not exhaustively capture the diversity of Indian women engaged in public-political life. Even so, it has yielded salient insights into the intersectional nature of violence faced by women online.

Sample distribution:

These filtering processes were used to narrow down the size of the sample to a total of 17 women under three categories, with a predetermined quota of at least 5 women from categories a., b. and c., to ensure an even distribution:

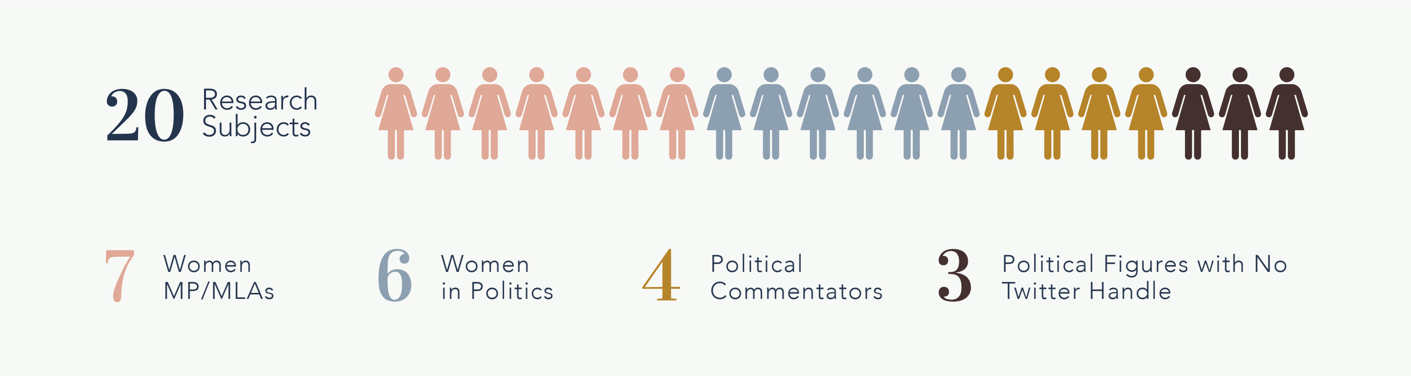
a. Women MPs/MLAs (Members of Parliament/Members of Legislative Assembly): Women who currently hold political office as members of legislature at either the Union or the state level.

b. Women in politics: Women who were formally involved in electoral party politics but are not currently in office.

c. Political commentators: Women who are not necessarily affiliated with any political party but are active contributors to public-political discourse online.

Besides individuals selected on the basis of their engagement rate on Twitter, we also wanted to account for women in political life who do not have a Twitter account but are nevertheless the subjects of misogynistic discussion and abuse by users on the platform. Thus, in addition to the three above-mentioned categories, we introduced a fourth, 'no Twitter handle' category with three women.

d. No Twitter Handle: Women MPs/MLAs and women politicians as defined under the first and second categories but without a Twitter account. The mentions directed at, or referring to, these women were collected by entering the women's names into the platform's search query.



Infographic 2: Sample Distribution

2.2 Data collection

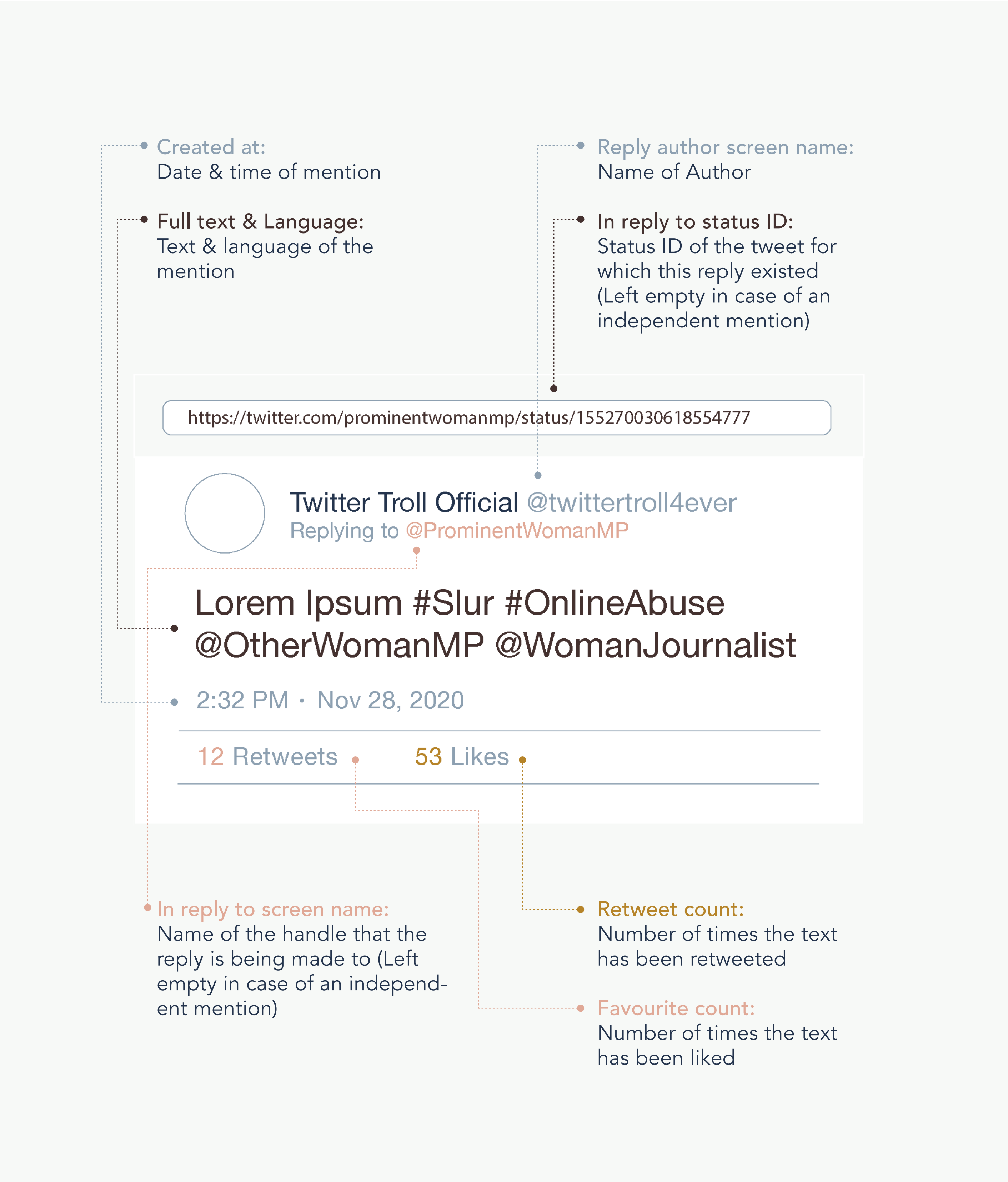
After the selection of the sample, mentions directed at the women were collected from their public Twitter profiles for a period of one week between **26 November to 3 December 2020.**

1	week of data collection	30,460	mentions collected annotated	8	data fields for each mention

This was done using a **python script with libraries like csv, tweepy, and pandas,** along with authorisation credentials provided through a Twitter developer account. This allowed us to search for interactions within a specified time period that included the subject's handle (for example, @JoeBiden).

For women belonging to the 'no Twitter handle' category, names were used as search terms (for example, Joe Biden or #JoeBiden). Throughout this report, we refer to each of these individual interactions as 'mentions,' which include replies, retweets, quote tweets, and independent mentions (meaning, not in reply to a woman in our sample).

Once the data was collected, the code allowed us to store it as an excel file type. In all, we **annotated 30,460 mentions**. Through the Twitter developer account, we can obtain a large number of fields for any particular mention. Based on our requirements, a data set with the fields outlined in Infographic 3 was populated.



Created at:
Date & time of mention

Full text & Language:
Text & language of the mention

Reply author screen name:
Name of Author

In reply to status ID:
Status ID of the tweet for which this reply existed (Left empty in case of an independent mention)

https://twitter.com/prominentwomanmp/status/155270030618554777

Twitter Troll Official @twittertroll4ever
Replying to @ProminentWomanMP

Lorem Ipsum #Slur #OnlineAbuse @OtherWomanMP @WomanJournalist

2:32 PM · Nov 28, 2020

12 Retweets   53 Likes

In reply to screen name:
Name of the handle that the reply is being made to (Left empty in case of an independent mention)

Retweet count:
Number of times the text has been retweeted

Favourite count:
Number of times the text has been liked

Infographic 3: List of fields populated for each mention

2.3 Annotation of mentions

After data collection, we developed a set of annotation guidelines to separate the hateful, abusive, and problematic tweets from the rest, and classify these into mutually exclusive categories of hateful or abusive speech. The annotation guidelines were developed on the basis of an inductive coding exercise. A total of 22 codes were defined for annotation to capture the nuances of hateful, abusive or problematic content (see Table 1). The annotation guidelines were used as a reference manual by three annotators who appraised and coded each of the mentions into the categories which defined the different kinds of hateful, abusive, and problematic speech.

Terminology	Definition
Derailing	Justifying female abuse, rejecting male privilege, attempting to disrupt the conversation in order to redirect the subject's opinions/views/perspectives to male-centred opinions/views/perspectives
Generic abuse	Using nasty/malevolent language with the intention to attack the subject because she is a woman
Sexualised slur	Using a pejorative like slut, whore, *presstitute* (a term used to attack women journalists), etc.
Sexist slur	Using a pejorative like bitch, *feminazi, witch, daayan* (witch in Hindi), etc.
Casteist slur	Using a casteist slur like *chamar, bhangi, kameeni* (these are derogatory terms used to refer to historically oppressed caste communities), etc.
Dehumanising insult	Comparing women to non-human beings
Exercising dominance	Asserting the superiority of men over women to naturalise gender inequality
Over-familiarity	Disrespecting the subject's personal boundaries, demonstrating creepiness
Stereotyping	Using a widely held but fixed and oversimplified image or notion of a woman/womanhood
Sexual harassment	Making sexual advances at or asking for sexual favours from the subject, inflicting harassment of a sexual nature

Terminology *contd.*	Definition *contd.*
Sexual objectification	Bullying based on physical characteristics such as skin colour, weight, body shape, looks; judging a woman's physical attractiveness based on patriarchal standards
Asexual objectification	Referring to the subject as an inanimate object
Intimidation	Replying with an intent to instigate fear
Threats of violence	Replying with an intent to physically assert power over the subject through threats of violence
Direct religious hate speech	Expressing hate towards the subject based on her religion
Religious stereotyping	Expressing hateful generalisation about the religion of the subject
Indirect religious hate speech	Attacking the subject perceived to be sympathetic to a minority religion
Casteist hate speech	Expressing hate towards the subject based on her/their caste
Delegitimising by othering	Delegitimising criticism on the basis of a narrow definition of nationality; asserting exclusive claim over a national/regional identity
Neutral	Text is neither abusive nor problematic
Other	Tweets that could not be easily categorised
Non-targeted abuse	Generic abusive statement not targeted at an identifiable individual

Table 1: Annotation guidelines

In order to ensure that there was agreement among the coders regarding the interpretation of the annotation guidelines, we conducted two inter-coder reliability (ICR) tests—one at the start of the annotation process, and another after one round of annotation by the coders. In ICR 1, we selected a random sample of 100 mentions, with proportionate representation from mentions directed at each of the women, and attained 40% agreement among the coders. In ICR 2, after the completion of one round of annotation, we created a sample of 264 mentions, with proportionate representation across the 22 hate speech codes, and attained 69% agreement. We considered this a fair-to-good agreement, considering the number of subcategories and the subtleties in these subcategories.

By the end of the annotation process, each of the hateful, abusive, and problematic mentions was independently appraised by the three coders. Disagreements about the classification of a particular mention were resolved through discussion among the coders. In cases where a consensus could still not be reached, a fourth research team member was asked to categorise the mention and resolve the disagreement.

Any specific mention could be categorised under a maximum of three mutually exclusive codes. This was done to account for the indeterminacy of the annotation guidelines and the fact that abusive tweets often do not fall neatly into any one category. For example, a tweet could simultaneously be categorised under 'religious hate speech,' 'a threat of violence,' and 'exercising dominance.' We annotated mentions in English, Hindi, Bengali, Marathi, Punjabi, Kannada, Gujarati, and a few in Tamil and Urdu. After the annotation process, we then divided the 19 codes of hateful, problematic or abusive speech into seven broader categories (see Table 2). This was done in order for us to be able to make more general claims about patterns of abusive speech, rather than relying on the highly specific definitions provided in the annotation guidelines.

1. Religious hate speech	Direct religious hate speech	Religious stereotyping	Indirect religious hate speech	
2. Abuse	Generic abuse	Asexual objectification	Dehumanising insult	
3. Misogyny	Derailing	Exercising dominance	Sexist slur	Stereotyping
4. Othering	Delegitimising by othering			
5. Sexualised references	Sexualised slur	Sexual harassment	Sexual objectification	Over-familiarity
6. Threat	Intimidation	Threats of violence		
7. Caste hate speech	Casteist slur	Casteist hate speech		

Table 2: Seven clubbed categories of hateful, problematic or abusive speech

2.4 Limitations

A primary limitation of our study is its small sample size of 20 women. To mention a few significant omissions, our sample does not adequately represent women politicians from northeast India or transgender women in public-political life. This is partially the result of our emphasis on a high engagement rate as a decisive criterion in the process of sample selection, which automatically excluded a large number of candidates whose locations may have otherwise been pertinent to a feminist socio-political analysis. Practical considerations such as time and resource constraints and proficiency of the team in Indian languages were other determining factors for this omission.

Another limitation of the study is the short time period in which data was collected. A duration of one week does not lend itself to understanding the growth and evolution of public-political discourse on social media platforms, which is an ongoing process. The need for longitudinal research that can examine the scale of misogyny online cannot be overemphasised. However, the fact that political discourse online is characterised by speed and ephemerality suggests that a study of this nature, which privileges episodic flashpoints, can provide vital insights into the morphology of societal relations and social power.[4]

Given that the digital divide is based on income, age, education, geographic location and other parameters also means that the cohort of women we studied may not be strictly representative of the wide spectrum of women in India's public-political life. While this does not dilute the significance of our findings about the sexism that characterises online publics, it does suggest the need to be alert to gendered exclusions and barriers that leave many women in public life without the ability to frame or respond to narratives in online discursive arenas that co-constitute the political terrain. The modest scale of our research project means that we cannot come to any authoritative or far-reaching claims about the quantum of gender-based violence on the platform. However, our research findings provide salient qualitative insights on the presence of discernible patterns of hate and abuse, and point towards potential directions for urgent remediative action and legal-institutional responses.

TRIGGER WARNING

Concerning the Privacy of the Research Subjects

Interspersed with the text of this report are verbatim reproductions of violent, abusive, misogynistic, offensive, and stereotyping mentions, which may be both disturbing and upsetting to read.
Please consider this a trigger warning.

Rather than relying solely on our own exposition,
we have reproduced these abusive mentions as we believe that an unsanitised presentation can convey the kinds of violent speech acts we reference, as well as the gravity of the broader issue of online gendered trolling more effectively. In the interest of protecting the women's privacy, their handles as well as any other personally identifiable information have been redacted.

03

Findings

This section presents some of the patterns of online misogyny that we were able to discern in our study. Some of these findings conform to internationally recognised patterns of abusive speech directed at women online, while others can be identified as the progression of culturally and historically specific forms of patriarchy.

We begin by outlining some of the most broad-level, generalisable findings (see Infographic 4) in section 3.1, and then proceed to highlight some of the findings in more minute detail, and with greater attention to cultural and political context.

Pervasiveness Of Misogynistic Speech

All women in our sample regardless of their political or ideological standpoints received some amount of abuse on the platform. No one was entirely spared.

Herd Aggression

Trolls tended to strategically target certain women and certain posts to exploit the affordances of virality and the algorithmic amplification of content.

“Light-hearted” Trolling

A majority of abusive messages directed at the women were of a supposedly milder variety of tongue-in-cheek jokes and remarks.

Intersectional Violence

Muslim women, political dissenters, and political commentators received an overwhelming majority of abusive messages.

Infographic 4: Broad patterns from data on the nature of misogyny

3.1 Broad patterns from data on the nature of misogyny

The broadest finding from our research is that all the women in our sample, regardless of whether they belong to opposition or ruling parties, and whether they are perceived to be dissenters or sympathetic to the current dispensation, faced some degree of abuse on the platform; none were entirely spared. What we may infer, therefore, is that trolls are located across the wide spectrum of ideological standpoints.

However, women 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 do not have any formal party affiliations were also at the receiving end of an inordinate amount of abuse. Figure 1 shows the distribution of problematic and abusive speech according to the clubbed categories in Table 2. Figure 2 then presents this same data, but is further broken up into four categories of women differently situated in the political field. The purpose of this disaggregation is to show how certain categories of women in public-political life are at the receiving end of a disproportionate amount of abuse.

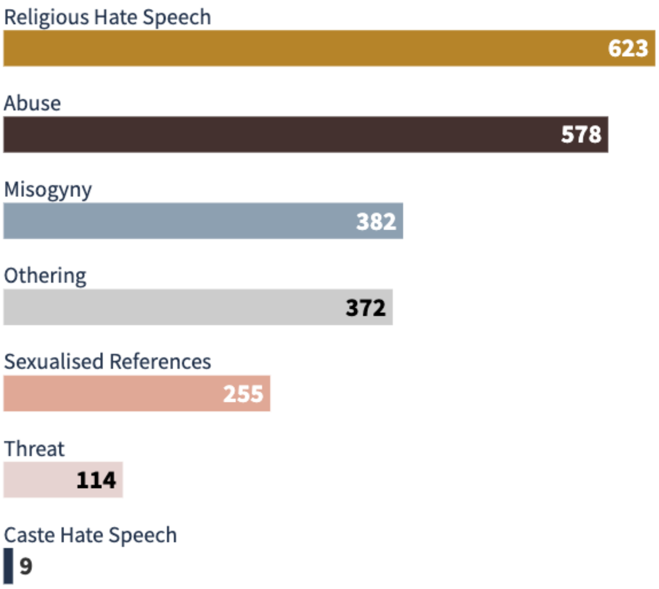


Figure 1: Instances of hate according to the seven clubbed categories

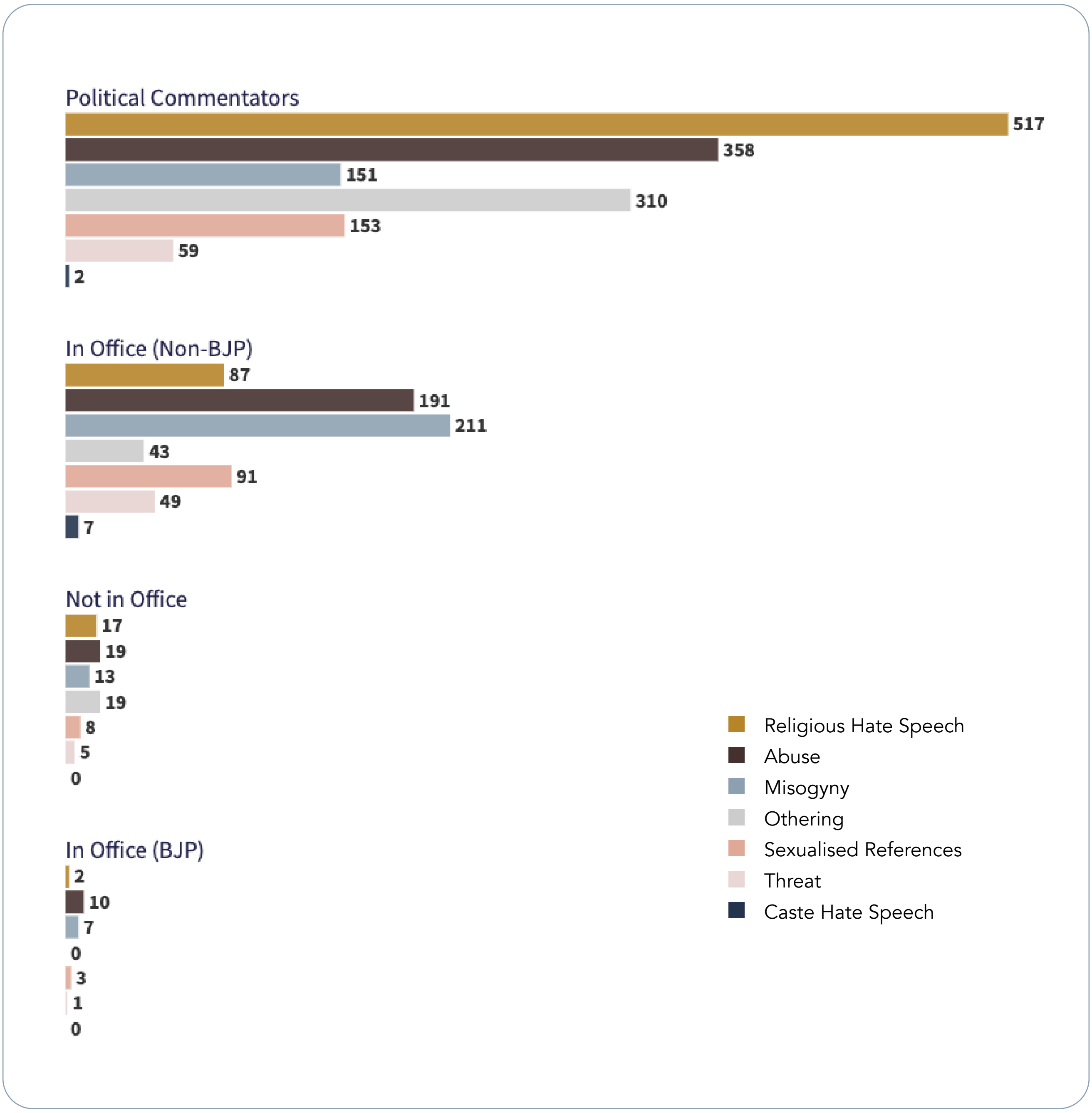


Figure 2: Distribution of each category of hateful, problematic or abusive speech across four categories of women in public-political life

Another counter-intuitive but crucial finding has to do with the most common categories of abusive gendered speech on the platform. Most abusive messages were not overtly grave, such as messages that wish death on women, threats of rape, etc., but rather, forms of trolling that are seemingly milder and in the nature of tongue-in-cheek jokes and remarks. We found this 'fun' culture of abuse to be rampant on the platform, especially through the sharing of misogynistic memes and wordplay.

We also found that the abusive speech directed at women in public life rarely had anything to do with their work or stated political positions. It invariably took the form of gendered attacks on their bodies or character. Rather than responding to or engaging with political positions (even with anger or abuse), trolls questioned women's credentials or trivialised their role in politics. This was done in a variety of ways, predominantly by insinuating that women were present in certain organisations or in certain capacities only to make up the numbers, and that the real authority lay with the men in their parties/family.

Another common tactic was derailing and whataboutery, where an entirely irrelevant incident or event in the life of the woman would be brought up with the sole intent of derailing the conversation. There were also numerous irrelevant comments made about women's appearances and how they were "all beauty and no brains". Another highly generalisable pattern we found had to do with how trolls targeted particular posts. Not every post by a public commentator/politician in our study received the same attention or response from trolls. Perhaps, partly as a way of establishing some kind of fraternal, masculinist solidarity, and partly as a way to exploit the algorithmic amplification on Twitter, we found the prevalence of a kind of herd aggression, where trolls, as homosocial, masculine communities, banded together to reply only to certain posts.

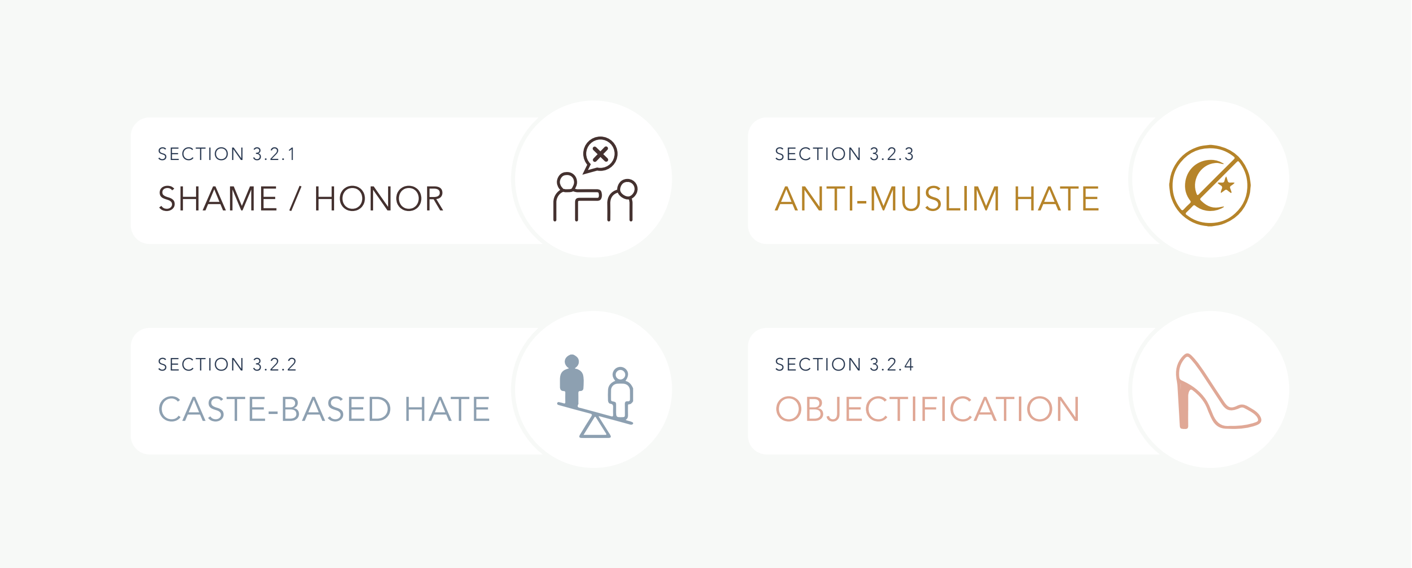
Our findings also indicate that trolls often tended to tag certain women together, as if to deride, warn, or intimidate all of them for their supposed allegiance, affiliation, or identity. Tags alluded to the ostensible membership in fabricated groups such as the *Tukde Tukde Gang*,[5] or other groupings such as Muslim women, Dalit politicians or women in politics who were previously part of the entertainment industry.

With the period of data collection limited to a week, this report obviously contains a skewed representation of political issues. Some of the most intensely debated issues at the time —that is, towards the end of November 2020— were the then ongoing farmers' protests, the untimely death of Bollywood actor Sushant Singh Rajput, Assembly elections in the state of West Bengal, and the enactment of anti-conversion laws in various states.[6] Given the constant churn of political content and the ways in which virality dictates engagement on Twitter, these divisive issues dwarfed any other political debates that may have arisen at the time.

3.2 The overarching subtext of Brahminical patriarchy

Corresponding to the broader range of behaviours of political and cultural regulation outlined above, the attendant misogynistic tropes that make up the patterns of abuse and censorship in the study carry a distinctive patriarchal flavour—that of Brahminical patriarchy.[7] Feminist sociologists and historians describe Brahminical patriarchy as the social-institutional order in which women's subordinate status, and their mobility, sexuality, choices, and desires are governed to maintain the supremacy and purity of a caste-based socially stratified order.

To uphold the absolutism of caste boundaries, Brahminical patriarchy deploys an interlocking set of norms, rules, and practices, critical among which is caste endogamy. The tropes of 'honour,' 'shame,' and 'respectability' are central to preserving the purity and preventing the pollution of the clan, and upholding them requires the gatekeeping of female sexuality. This endeavour turns women into flagbearers for the household and its social standing, and repositories of community pride.[8] Denied both autonomy and agency, and circumscribed to their roles in the domestic sphere, women in Brahminical patriarchy, thus, become the veritable objects of male entitlement, domination, and control.



SECTION 3.2.1
SHAME / HONOR

SECTION 3.2.3
ANTI-MUSLIM HATE

SECTION 3.2.2
CASTE-BASED HATE

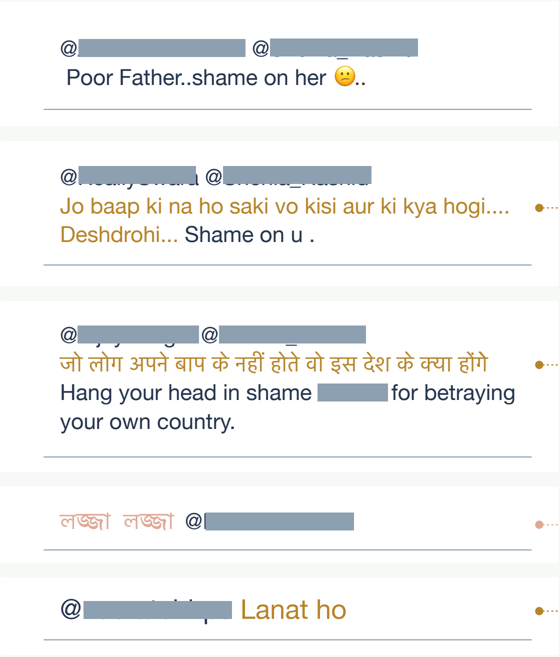
SECTION 3.2.4
OBJECTIFICATION

Infographic 5: The overarching subtext of Brahminical patriarchy. ***Click within the graphic to go to the relevant section.***

3.2.1 Shame/Honour

This preoccupation with shame and honour was highly prevalent in the mentions we studied. Many of the abusive tweets that we annotated contained some mention of the words "shame," "shameless," "honour," "*laanat*," or "*sharm*" (the last two words mean 'shame,' in Hindi). "Shame on you" or "Hang your head in shame" and their equivalent in Indian languages was a common refrain, often used to convey the message that the woman had not only damaged her own reputation (for whatever perceived indiscretion), but had brought shame upon her husband, father or community.

SNAPSHOTS



@[redacted] @[redacted]
Poor Father..shame on her 😕..

@[redacted] @[redacted]
Jo baap ki na ho saki vo kisi aur ki kya hogi.... Deshdrohi... Shame on u .

(Hindi) If she can't even belong to her own father, how can she belong to anyone else... Traitor...

@[redacted] @[redacted]
जो लोग अपने बाप के नहीं होते वो इस देश के क्या होंगे
Hang your head in shame [redacted] for betraying your own country.

(Hindi) If someone doesn't belong to their father, how can they belong to their nation?

লজ্জা লজ্জা @[redacted]

(Bengali) For shame

@[redacted] Lanat ho

(Hindi) Shame on you

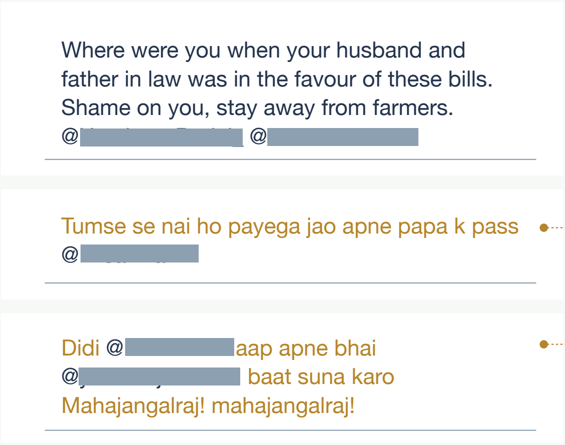
Writing about the horrific instances of sexual abuse inflicted on women on either side of the India-Pakistan border during the 1947 partition, Das speaks about how values of honour and shame are inscribed on the bodies of women, "doubly articulated in the domains of kinship and politics."[9] In the aftermath of the partition that created a Hindu majority India, and Muslim majority Pakistan, the two countries entered into an agreement to mutually "restore" their women to their home countries. The Indian government enacted a law titled "Abducted Persons (Recovery and Restoration) Act, 1949"[10] to "recover" Hindu and Sikh abducted women in Pakistan, and "restore" them to their families in India.

Similarly, Muslim women left behind in India were to be "restored" to their migrated families in Pakistan. The sexual violence inflicted on women is not only a stand-in for shaming the community being othered, it is also a proxy for the 'wounded honour' of the nation. In the years since the 1947 partition, and now on social media, the 'dishonouring' of women continues to be a means of othering enacted through the puritanical logic of Brahminical patriarchy.

We found this sort of othering and hypernationalistic rhetoric targeted most often at Muslim women who are seen to represent an imminent threat to the moral purity of the nation. In a recent controversy on the subject of religious conversion and interfaith marriage between a Sikh woman and a Muslim man, which allegedly involved forced conversion,[11] a Sikh man tweeted: "No. Girls will not marry Muslims. Live with it. Deal with it. This is the line we have drawn at a community level."

This is further proof that women in the South Asian context are not seen as individuals in their own right, but rather, as subjects of male authority and control; their bodies are marked as repositories of community values, and become sites where contestations of family, community, and national honour play out. Very often, the women in our sample received mentions that reduced their contribution in politics to that of supporting roles to the men in their families. Women who come from political families constantly received comments about how they are supposedly diminishing the name and reputation of their families.

SNAPSHOTS



Where were you when your husband and father in law was in the favour of these bills. Shame on you, stay away from farmers.
@[redacted] @[redacted]

Tumse se nai ho payega jao apne papa k pass
@[redacted]

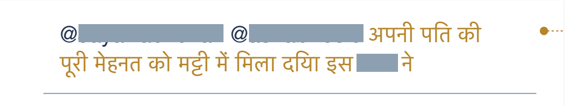
Didi @[redacted]aap apne bhai
@[redacted] baat suna karo
Mahajangalraj! mahajangalraj!

(Hindi) You won't be able to handle this, go to your daddy.

(Hindi) Sister, please listen to your brother's advice on the issue.
Lawless land! Jungle Raj!

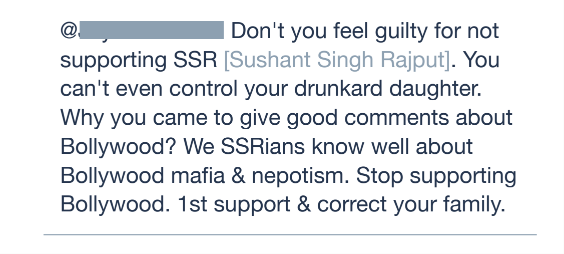
Trolls made similar remarks about women politicians whose husbands or other male family members were prominent in their own fields. A common retort to any assertions of a political position/opinion was that the woman had dragged her husband's [or male family member's] name through the dirt. Similarly, an irrelevant episode from the woman's past or personal life was often brought up to discredit her political claims. Putting women in their 'rightful' place in the domestic sphere by raking up their private lives in public view was another common pattern we observed.

SNAPSHOTS



@[redacted] @[redacted] अपनी पति की पूरी मेहनत को मट्टी में मिला दिया इस [redacted] ने

(Hindi) This woman has driven her husband's good name into the dirt.



@[redacted] Don't you feel guilty for not supporting SSR [Sushant Singh Rajput]. You can't even control your drunkard daughter. Why you came to give good comments about Bollywood? We SSRians know well about Bollywood mafia & nepotism. Stop supporting Bollywood. 1st support & correct your family.

3.2.2 Caste-based hate

A second aspect that characterises misogyny in the Indian online public sphere is the widespread prevalence of caste-based/caste-directed hate. Although we found indirect references to Brahminical notions of purity and honour in mentions directed at all the women in our sample, the abusive mentions that targeted women from marginalised caste groups in our sample took on a distinctly different hue.

The theory of Brahminical patriarchy, and its underlying concept of "graded inequality"[12] shows how Dalit women in particular, are susceptible to violence that is predicated on both their caste and gender identities. Highlighting the specific ways in which Dalit women are abused online, Kiruba Munusamy writes,

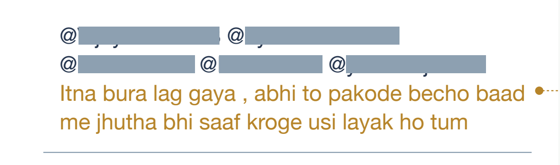
> *"When women are generally threatened with rapes and slut-shaming, outcaste women are insulted as unworthy or too ugly to rape, or labelled as being a slut is hereditary and predominantly because of*

> *being born in the untouchable caste. When privileged women are criticised for their personal choices, outcaste women are criticised for their choices like food that comes from their cultural background."*[13]

Our findings point to the need to recognise the crucial role played by caste as an axis of social power in structuring the nature of online misogyny. The abuse directed at the Dalit women in our sample took a range of forms.

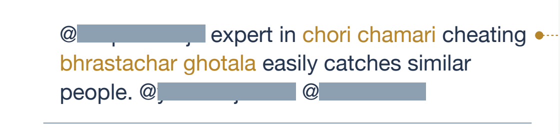
We found the use of casteist slurs such as *chamar, kameeni, bhand,*[14] as well as other direct or unveiled forms of casteism such as references to *jootha*—a practice of untouchability[15]—or calling a Dalit politician *neech*, a Hindi word that is meant to signify one's lowly social status, with casteist connotations.[16]

SNAPSHOTS



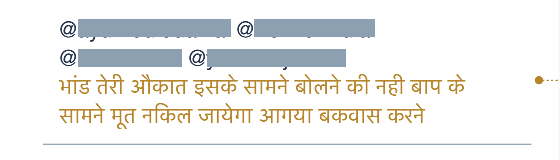
@[redacted] @[redacted]
@[redacted] @[redacted] @[redacted]
Itna bura lag gaya , abhi to pakode becho baad me jhutha bhi saaf kroge usi layak ho tum

(Hindi) Go and sell pakodas and clean up afterwards, that is all you are worthy of.



@[redacted] expert in chori chamari cheating bhrastachar ghotala easily catches similar people. @[redacted] @[redacted]

(Hindi) You are an expert in thievery, corruption and scamming, and easily catch people of a similar character.



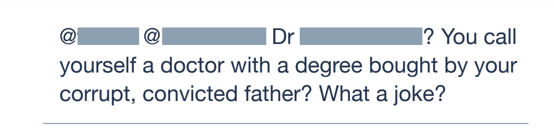
@[redacted] @[redacted]
@[redacted] @[redacted]
भांड तेरी औकात इसके सामने बोलने की नही बाप के सामने मूत नकिल जायेगा आगया बकवास करने

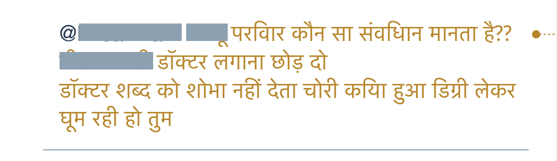
(Hindi) *Bhand*, you are not worthy of speaking in front of him, you come here and talk nonsense

Our findings suggest, however, that casteist stereotyping and insinuations of casteism are more common than direct casteist attacks. One common and familiar stereotype was to call into question the merit of the women in our sample through the use of words such as *aukaat* (loosely translates to status/ability), *ghotalebaaz* (scamster), duplicate certificate *wali*, uneducated, *chor* (thief).

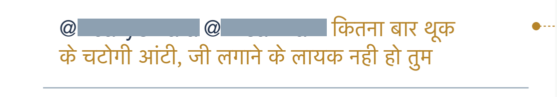
Allegations of corruption were also levelled disproportionately against Dalit women politicians, with aspersions cast on their professional integrity.

SNAPSHOTS


@[redacted] @[redacted] Dr [redacted]? You call yourself a doctor with a degree bought by your corrupt, convicted father? What a joke?

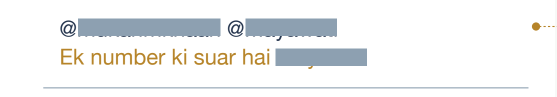

@[redacted] [redacted] परविार कौन सा संवधिान मानता है??
[redacted] डॉक्टर लगाना छोड़ दो
डॉक्टर शब्द को शोभा नहीं देता चोरी कयिा हुआ डिग्री लेकर घूम रही हो तुम

(Hindi) What constitution does your family abide by? Please do not become a doctor, you are insulting the profession by having acquired a degree through cheating.


@[redacted] @[redacted] कितना बार थूक के चटोगी आंटी, जी लगाने के लायक नही हो तुम

(Hindi) How many times will you spit out and lap it up aunty? It isn't worth engaging with you.

As mentioned in section 3.1, we found that trolls often tagged people who were assumed to hold similar political beliefs. This was the case with Dalit politicians and Ambedkarites as well. Many of them were often tagged and abused together, and there were even calls for "people like them" to be "encountered."[17] We also found multiple instances of dehumanising insults directed at Dalit women politicians. They were called *suar* (pig), which not only carries casteist connotations of uncleanness, but also ties in with notions of traditional caste-based occupations, suggesting that the women in question can only belong in these 'filthy' and inhumane occupations.


@[redacted] @[redacted]
Ek number ki suar hai [redacted]

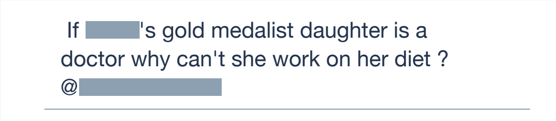
(Hindi) She is a pig of the first order.

Even though instances of caste-based hate speech in our sample were far fewer than other categories such as religion-based hate speech or othering speech, our findings show that the vice of casteism does emerge as a definite problem on social media. Indeed, there is ample evidence of the extensive use of caste-based hate speech on Facebook as well.[18]

It is possible that the deliberate erasure and invisibilisation of caste identity is reflected in the online behaviour of trolls, given the current misplaced tendency to view caste as a thing of the past. Based on the disproportionate amount of abusive messages directed at women in our sample who are perceived to be on the left of the political spectrum, we can safely assume that the majority of trolls we encountered are aligned with the Hindu right. The claim that caste oppression has been transcended is certainly not borne out by any evidence. As has been pointed out by scholars such as Satish Deshpande, it is a calculated political strategy of the ruling party to secure electoral success by winning over a broader range of caste constituencies toward a unified Hindutva politics.[19]

Yet another reason that caste-based hate was less evident in our sample, may also be connected to the methodology adopted for annotation and its inadequacies in unpeeling the subtle ways by which gendered hate manifests in socio-cultural discourse. Although our annotation guidelines accounted for the indeterminacy of each category by allowing for any particular mention to be classified into upto three different codes, surfacing caste-based attacks was not always easy. As mentioned earlier, caste-based domination on social media platforms is often based on insinuation, rather than direct attacks.

Applying an intersectional lens, multiple ideologies of oppression operate cumulatively and concurrently to produce a specific experience of subordination. A mention which is potentially categorisable as 'exercising dominance' or 'over-familiarity,' when directed at a Dalit woman, may take on the colour of casteist speech. However, the imperfect process of classifying a particular mention into categories of problematic or abusive speech in accordance with the annotation guidelines was not always able to capture this analytical depth. Consider the following example:



If [redacted]'s gold medalist daughter is a doctor why can't she work on her diet ?
@[redacted]

We classified this tweet under 'sexual objectification,' as it is a comment about the woman's appearance. But it could also be read as an attempt to poke fun at her for being the topper, insinuating that as a 'lower-caste' woman, she is undeserving of the gold medal.

The lexical meaning of a word arises in relation to a broader discourse—the context and the text.[20] Any annotation schema for a medium like Twitter—even if built inductively and worked on by an involved team of human coders, such as in our study—may not be able to capture fully or accurately the social meaning of a word, as well as the values and ideas underlying such meaning. The challenge of unpacking semantic discourse on social media is greater when it comes to deeply embedded social hierarchies of sexism or casteism that are normalised in everyday parlance. Yet, as the textual syntactic structures of social media become part of the very process of producing and understanding meaning, they not only begin to structure attitudes, but also the context.

Platforms like Twitter, thus, become powerful structuring tools, the very context that reinforces marginal gendered positions in ways that may be difficult to decipher. We delve deeper into the ways in which social media platforms fundamentally alter the underlying conditions of social interaction and meaning-making in section 4.1.3.

The fact that caste-based attacks on women in public-political life are often subtle and based on insinuations prompts us to think about how abuse can take on culturally rooted forms that may seem innocuous if decontextualised.[21] However, within the social mores that define relationships of power (as in Brahminical patriarchy), these forms of abuse are entrenched in the hegemonic order, legitimising oppression. This has implications for both how the law, as well as platforms should respond in their content governance policies, and how social media algorithms should be used to filter objectionable content.

The findings of the study also suggest that western theories of hate speech may not be able to adequately capture the nature of caste-based abuse in India, and that hate speech may not be equivalent to discriminatory speech.[22] These are pertinent considerations in developing a context-specific legal-institutional response to such speech.

3.2.3 Anti-Muslim hate

An overwhelming majority of hateful, abusive, and misogynistic mentions were directed at the Muslim women in our sample. Anti-Muslim rhetoric and violence have been steeply on the rise over the past decade.

While historical fault lines in the relations between Muslims and majority Hindus run deep, divisive rhetoric and on-ground polarisation have reached a fever pitch since the widespread adoption of social media in India.[23] The rise of right-wing populism and the ideologies of the *Sangh Parivar*[24] have normalised diatribe against Muslims online.[25]

Our research provides a glimpse into this thriving online ecosystem of Islamophobic propaganda. It is difficult to overstate the sheer volume of abusive and hateful mentions directed at politically prominent Muslim women on Twitter.
Broadly speaking, our findings show that most of these mentions fall into two categories: The first is the use of hypernationalistic and othering speech that attacks Muslim women on the basis of a narrow definition of nationality or by asserting an exclusive claim over national identity. This includes the use of terms such as traitor, *gaddar* (traitor), *Tukde Tukde Gang*.[26] This seems to follow a globally recognised pattern of targeting minority groups through accusations of collusion with foreign intelligence agencies and treason, and through the use of violent and threatening speech as a means of intimidation.[27]

SNAPSHOTS

@[redacted] You have lavish lifestyle and a failed fake journo cannot afford this. Who funds you?Are you ISI agent?

@[redacted] i think somebudy will burn your house one day

Is [redacted] @[redacted] pregnant ? If not - why is she being allowed to roam around free? Such anti-nationals have only once place for them - Jail or Pakistan. Take your pick.

@[redacted] @[redacted]
Nikal bsdk pakistan chali ja R@nd

(Hindi) Get out, go to Pakistan, whore.

@[redacted] @[redacted] @[redacted]
@[redacted] [redacted]....Tumhara number bhi aanewala hai.. kagaj dhoond lo jaldi....

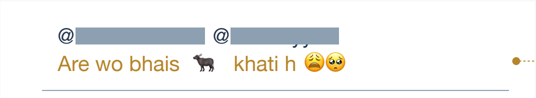
(Hindi) Your turn is coming soon too, better find your papers [documentation] quickly.

@[redacted] Tusssi gaddar ho madam ji...

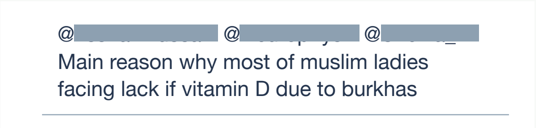
(Hindi) Dear ma’am, you are a traitor.

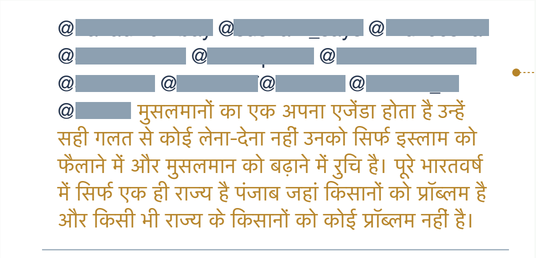
A second category of mentions either expressed stereotypical, hateful, and propaganda-informed generalisations about Islam, or exhibited hate towards the woman based on her religion. In this category, we found a large volume of tweets that were meant to engender fear among the Muslim community, as well as engineer false anxieties about the community posing a danger to the integrity of the nation. The following mentions are indicative of the troubling extent to which anti-Muslim fear speech[28] has spread online.

SNAPSHOTS

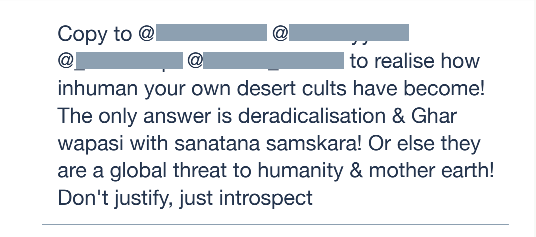

@[redacted] @[redacted]
Are wo bhais 🐃 khati h 😩🥺

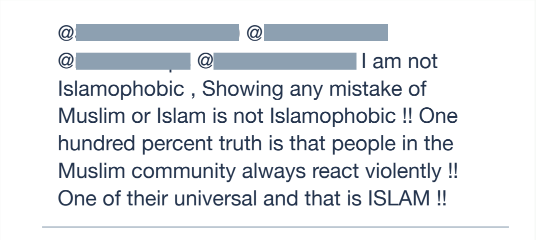
(Hindi) Aye, she eats beef.

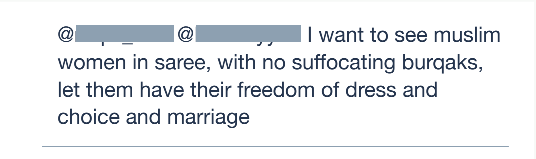

@[redacted] @[redacted] @[redacted]
Main reason why most of muslim ladies facing lack if vitamin D due to burkhas

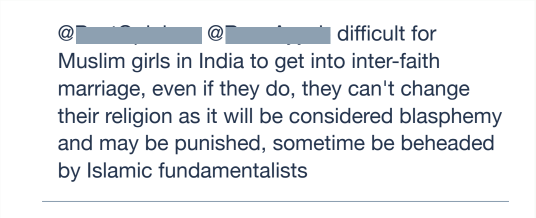

@[redacted] @[redacted] @[redacted] @[redacted] @[redacted] @[redacted] @[redacted] @[redacted] @[redacted] @[redacted] @[redacted] मुसलमानों का एक अपना एजेंडा होता है उन्हें सही गलत से कोई लेना-देना नहीं उनको सिर्फ इस्लाम को फैलाने में और मुसलमान को बढ़ाने में रुचि है। पूरे भारतवर्ष में सिर्फ एक ही राज्य है पंजाब जहां किसानों को प्रॉब्लम है और किसी भी राज्य के किसानों को कोई प्रॉब्लम नहीं है।

(Hindi) Muslims have their own agenda, they do not care about right or wrong. All they are interested in is in spreading Islam. In all of India, Punjab is the only state where farmers are facing problems, nowhere else.


Copy to @[redacted] @[redacted] @[redacted] @[redacted] to realise how inhuman your own desert cults have become! The only answer is deradicalisation & Ghar wapasi with sanatana samskara! Or else they are a global threat to humanity & mother earth! Don't justify, just introspect

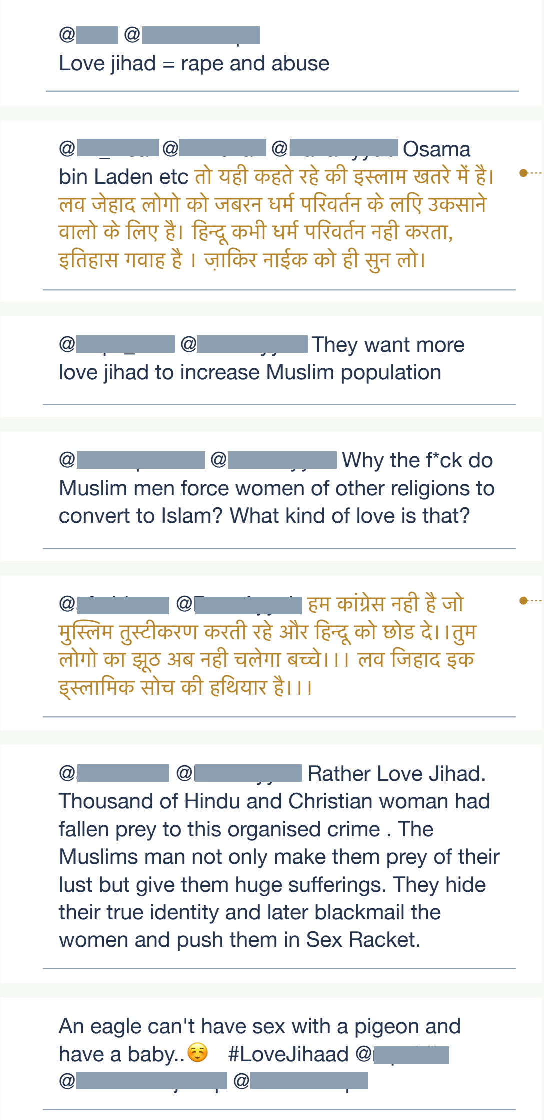

@[redacted] @[redacted] @[redacted] @[redacted] I am not Islamophobic , Showing any mistake of Muslim or Islam is not Islamophobic !! One hundred percent truth is that people in the Muslim community always react violently !! One of their universal and that is ISLAM !!


@[redacted] @[redacted] I want to see muslim women in saree, with no suffocating burqaks, let them have their freedom of dress and choice and marriage


@[redacted] @[redacted] difficult for Muslim girls in India to get into inter-faith marriage, even if they do, they can't change their religion as it will be considered blasphemy and may be punished, sometime be beheaded by Islamic fundamentalists

We also found several utterances of the word *jihad*, a deeply layered concept in Islamic thought,[29] corrupted into an oversimplified idea through which it has come to be understood in the popular imagination. The term "*love jihad*"—the misplaced idea that Muslim men target and seduce Hindu women for conversion to Islam as part of a broader 'war' by Muslims against India—is widely deployed by trolls. This, along with other spin-off terms such as "*corona jihad,*" "*narcotics jihad,*" etc., signify deep-seated Islamophobia.

SNAPSHOTS



@[redacted] @[redacted]
Love jihad = rape and abuse

@[redacted] @[redacted] @[redacted] Osama bin Laden etc तो यही कहते रहे की इस्लाम खतरे में है। लव जेहाद लोगो को जबरन धर्म परिवर्तन के लिए उकसाने वालो के लिए है। हिन्दू कभी धर्म परिवर्तन नही करता, इतिहास गवाह है । ज़ाकिर नाईक को ही सुन लो।

(Hindi) So everyone says that Islam is under threat. But love Jihad is a tool for forced religious conversion. Hindus have never carried out such conversions, history bears witness to this. Just listen to Zakir Naik.

@[redacted] @[redacted] They want more love jihad to increase Muslim population

@[redacted] @[redacted] Why the f*ck do Muslim men force women of other religions to convert to Islam? What kind of love is that?

@[redacted] @[redacted] हम कांग्रेस नही है जो मुस्लिम तुस्टीकरण करती रहे और हिन्दू को छोड दे।।तुम लोगो का झूठ अब नही चलेगा बच्चे।।। लव जिहाद इक इस्लामिक सोच की हथियार है।।।

(Hindi) We are not like the Congress who do Muslim appeasement and pay no attention to Hindus. Your lies will not work any longer. Love Jihad is a weapon of this Islamic thinking.

@[redacted] @[redacted] Rather Love Jihad. Thousand of Hindu and Christian woman had fallen prey to this organised crime . The Muslims man not only make them prey of their lust but give them huge sufferings. They hide their true identity and later blackmail the women and push them in Sex Racket.

An eagle can't have sex with a pigeon and have a baby..☺️ #LoveJihaad @[redacted]
@[redacted] @[redacted]

In addition to *jihadi* (a person involved in *Jihad*), other terms that we found being used to refer to Muslim women were *bibi, begum, halala.* Again, these words are not abusive in themselves, and are often used to refer to Muslim women (*bibi* and *begum*) or to specific religious practices (*halala*), but they were used by trolls as a means of othering and to mark out the religion of the women they were attacking. The week in which we collected data for this study coincided with the introduction of anti-conversion legislation, dubbed as "*anti-Love Jihad*" laws, in certain states.[30]

Our research shows the prevalence of inauthentic user behaviour in the numerous cookie-cutter tweets on the subject of "*love jihad,*" which is evidence of organised campaigns to spread conspiracy theories about Muslim men supposedly entrapping Hindu women.[31] In the period since data collection, orchestrated right-wing campaigns against Muslim women have become routinised on social media, with despicable attacks through apps/channels such as *Sulli Deals and Bulli Bai*[32] on which pictures of Muslim women were hosted without their permission and put up for "auction."

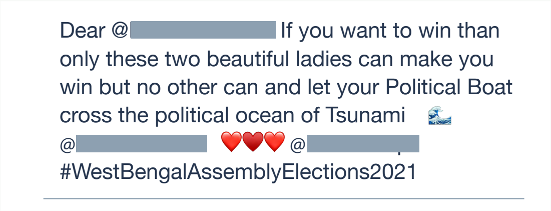
To be sure, most of the abusive tweets directed at Muslim women in our sample fell into the category of 'religious hate speech' and 'othering.' Besides this category, Muslim women who were part of the study received the most abuse in other categories such as 'generic abuse,' 'sexualised references,' and 'threats of violence.' This echoes the point made in section 3.2.2 about the need to be attentive to specific intersectional social locations.

A Muslim woman's experience of misogyny is not simply the sum of sexism and religious discrimination; it lies in the specific ways in which these factors are fundamentally constitutive of subjecthood and social identity. For example, Muslim women from our sample, who also happen to be dissenters or outspoken critics of the ruling dispensation, received abuse for being a dissenter and Muslim, but these were not the only determining factors. Similarly, Muslim politicians in our sample received more comments about their appearance than their Hindu counterparts.

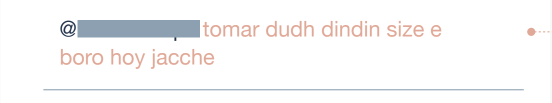
3.2.4 Objectification

An exceedingly common trope was to view women solely as the objects of male desire, reducing them to their bodies. This tendency took shape in a few different ways such as posting comments that hypersexualised women and passing unsolicited comments about their appearance.

SNAPSHOTS

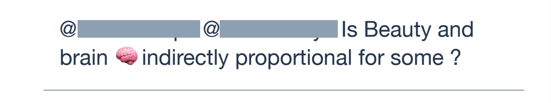


Dear @[redacted] If you want to win than only these two beautiful ladies can make you win but no other can and let your Political Boat cross the political ocean of Tsunami 🌊
@[redacted] ❤️❤️❤️ @[redacted]
#WestBengalAssemblyElections2021

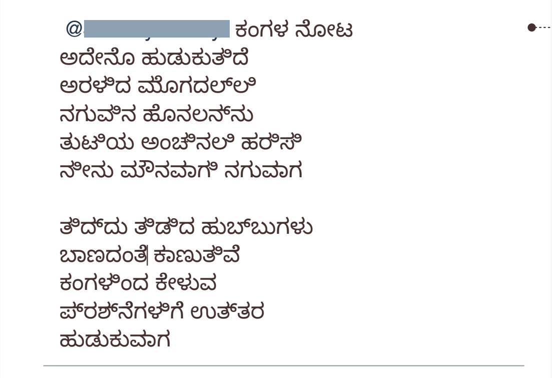


@[redacted] tomar dudh dindin size e boro hoy jacche

(Bengali) Your breasts are getting bigger and bigger each day.



@[redacted] @[redacted] Is Beauty and brain 🧠 indirectly proportional for some ?



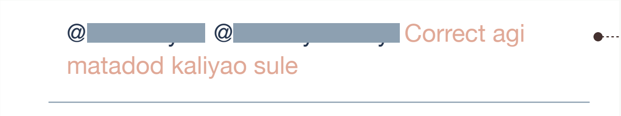
@[redacted] ಕಂಗಳ ನೋಟ
ಅದೇನೊ ಹುಡುಕುತ್ತಿದೆ
ಅರಳಿದ ಮೊಗದಲ್ಲಿ
ನಗುವಿನ ಹೊನಲನ್ನು
ತುಟಿಯ ಅಂಚಿನಲ್ಲಿ ಹರಿಸಿ
ನೀನು ಮೌನವಾಗಿ ನಗುವಾಗ

ತಿದ್ದು ತಿಡಿದ ಹುಬ್ಬುಗಳು
ಬಾಣದಂತೆ| ಕಾಣುತ್ತಿವೆ
ಕಂಗಳಿಂದ ಕೇಳುವ
ಪ್ರಶ್ನೆಗಳಿಗೆ ಉತ್ತರ
ಹುಡುಕುವಾಗ

(Kannada) Your searching eyes
As if seeking something
Upon the blooming countenance
Is your stream of laughter
Poured along the edges of lips
When you laugh silently

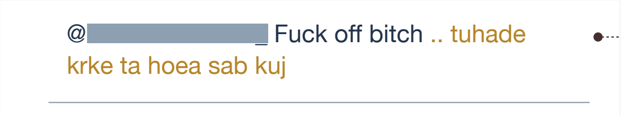
Your embellished eyebrows
Resembles that of a bow
While searching
Answers to the questions
Asked through the eyes

In addition to these comments that sexualised women and addressed them in needlessly over-familiar terms, women were also censured and attacked through the use of invectives and sexualised slurs such as *randi* (prostitute in Hindi), bitch, *sule* (prostitute in Kannada).



@[redacted] @[redacted] Correct agi matadod kaliyao sule

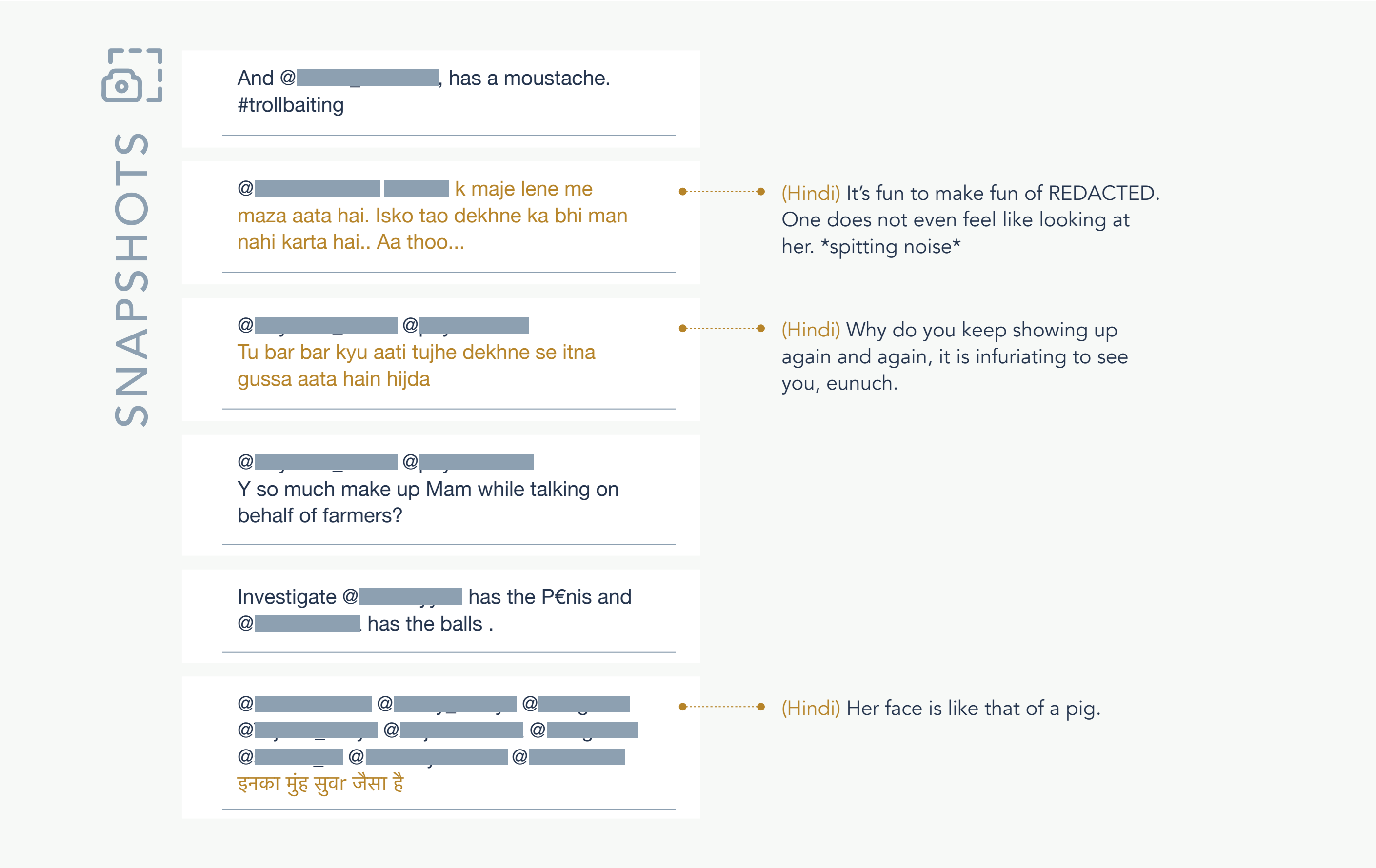
(Kannada) Learn how to talk properly, slut.



@[redacted] Fuck off bitch .. tuhade krke ta hoea sab kuj

(Punjabi) ...all this is your doing.

By the same token, women considered 'undesirable' by trolls, received misogynistic hate that took the form of comments expressing disgust or repulsion about their appearance, often through the use of terms such as *hijda* (eunuch), *chudail* (witch), *komolika* (the name of an iconic 'vamp' villain[33] in a Hindi TV serial), and *suar* (pig).



SNAPSHOTS

And @[redacted], has a moustache.
#trollbaiting

@[redacted] [redacted] k maje lene me maza aata hai. Isko tao dekhne ka bhi man nahi karta hai.. Aa thoo...

(Hindi) It's fun to make fun of REDACTED. One does not even feel like looking at her. *spitting noise*

@[redacted] @[redacted]
Tu bar bar kyu aati tujhe dekhne se itna gussa aata hain hijda

(Hindi) Why do you keep showing up again and again, it is infuriating to see you, eunuch.

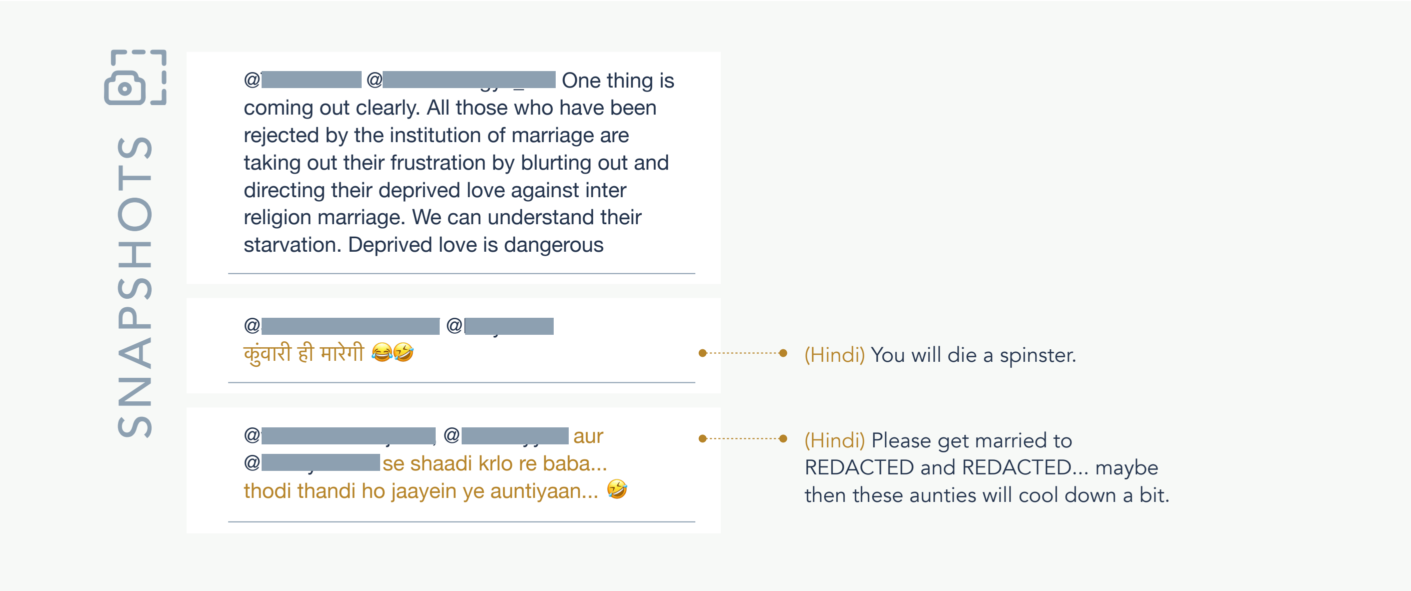
@[redacted] @[redacted]
Y so much make up Mam while talking on behalf of farmers?

Investigate @[redacted] has the P€nis and @[redacted] has the balls .

@[redacted] @[redacted] @[redacted]
@[redacted] @[redacted] @[redacted]
@[redacted] @[redacted] @[redacted]
इनका मुंह सुवर जैसा है

(Hindi) Her face is like that of a pig.

Another strain of objectifying mentions relates to a form of trolling that sought to police women's sexuality and sexual expression. We found instances of women being abused on either extreme of this spectrum. Some women were desexualised, called sex-deprived and referred to as aunty, *chachi* (aunt in Hindi) or *pishi* (aunt in Bengali) while others were shamed for any kind of sexual expression.



SNAPSHOTS

@[redacted] @[redacted] One thing is coming out clearly. All those who have been rejected by the institution of marriage are taking out their frustration by blurting out and directing their deprived love against inter religion marriage. We can understand their starvation. Deprived love is dangerous

@[redacted] @[redacted]
कुंवारी ही मारेगी 😂🤣

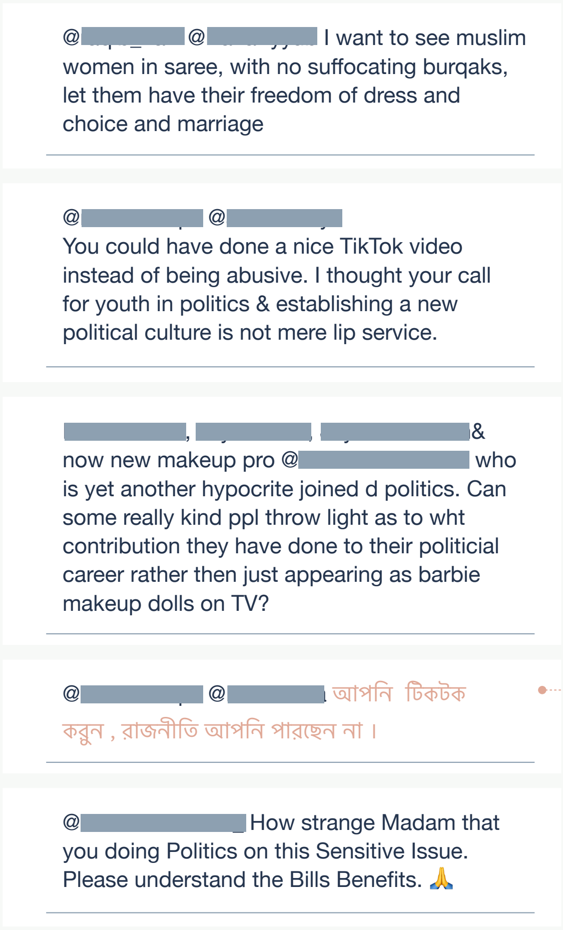
(Hindi) You will die a spinster.

@[redacted] @[redacted] aur
@[redacted] se shaadi krlo re baba...
thodi thandi ho jaayein ye auntiyaan... 🤣

(Hindi) Please get married to REDACTED and REDACTED... maybe then these aunties will cool down a bit.

Even in cases where trolls did not explicitly use foul and demeaning language, they addressed women filtered through a rigid set of patriarchal gender norms. This included gendered and religious stereotyping, assertions that women should stick to the things they are good at, and insinuations that politics is no place for women.

SNAPSHOTS



@[redacted] @[redacted] I want to see muslim women in saree, with no suffocating burqaks, let them have their freedom of dress and choice and marriage

@[redacted] @[redacted]
You could have done a nice TikTok video instead of being abusive. I thought your call for youth in politics & establishing a new political culture is not mere lip service.

[redacted], [redacted], [redacted]& now new makeup pro @[redacted] who is yet another hypocrite joined d politics. Can some really kind ppl throw light as to wht contribution they have done to their politicial career rather then just appearing as barbie makeup dolls on TV?

@[redacted] @[redacted] আপনি টিকটক করুন , রাজনীতি আপনি পারছেন না ।

(Bengali) You should stick to making Tik Tok videos, you won't be able to do politics.

@[redacted] How strange Madam that you doing Politics on this Sensitive Issue. Please understand the Bills Benefits. 🙏

These are only a handful of the entire gamut of mentions that we could have reproduced to show the ways in which women are objectified on Twitter. These instantiations speak to a widespread culture of impunity in which trolls feel entitled and emboldened to make comments such as these without fear of repercussions.

What they have in common is a view of womanhood rooted in a culturally specific patriarchal imagination, which either idealises women and places them on a pedestal or considers them only worthy of disdain. Every once in a while, this kind of content gains traction and is widely discussed by online commentators.

These events are often precipitated by men who feel compelled to police women and proffer their unsolicited opinion on what constitutes appropriate Indian womanhood.[34] But as evidenced by our data from just one ordinary week on the platform, this kind of objectifying rhetoric is flourishing online.[35]

By studying abusive speech directed at women in public-political life, we are uniquely positioned to see the links between culturally specific forms of gendered regulation (that we have elaborated under the conceptual umbrella of Brahminical patriarchy), and ideological constructions of the nation which have come to gain prominence in the online public sphere.

Much of the gendered violence that we observed was couched in the language of a muscular and exclusionary Hindu nationalism, and the persistent threat of male control over women's sexual autonomy cut across all experiences of misogyny, without exception. Our findings, therefore, bring into sharp relief the intimate linkages between the construction of the nation as an imagined community, and the appropriate role of gendered subjects therein.

Analysis

04

The first part of this section attempts to understand online misogyny and the ways in which the affordances of social media platforms engender new speech practices and distinct forms of sociality through two convergent inquiries. 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 kinds of vulnerabilities for women? These questions take us some way toward recognising the deep linkages between the cultural specificity of abusive speech practices and the emergent infrastructural context in which such speech is normalised.

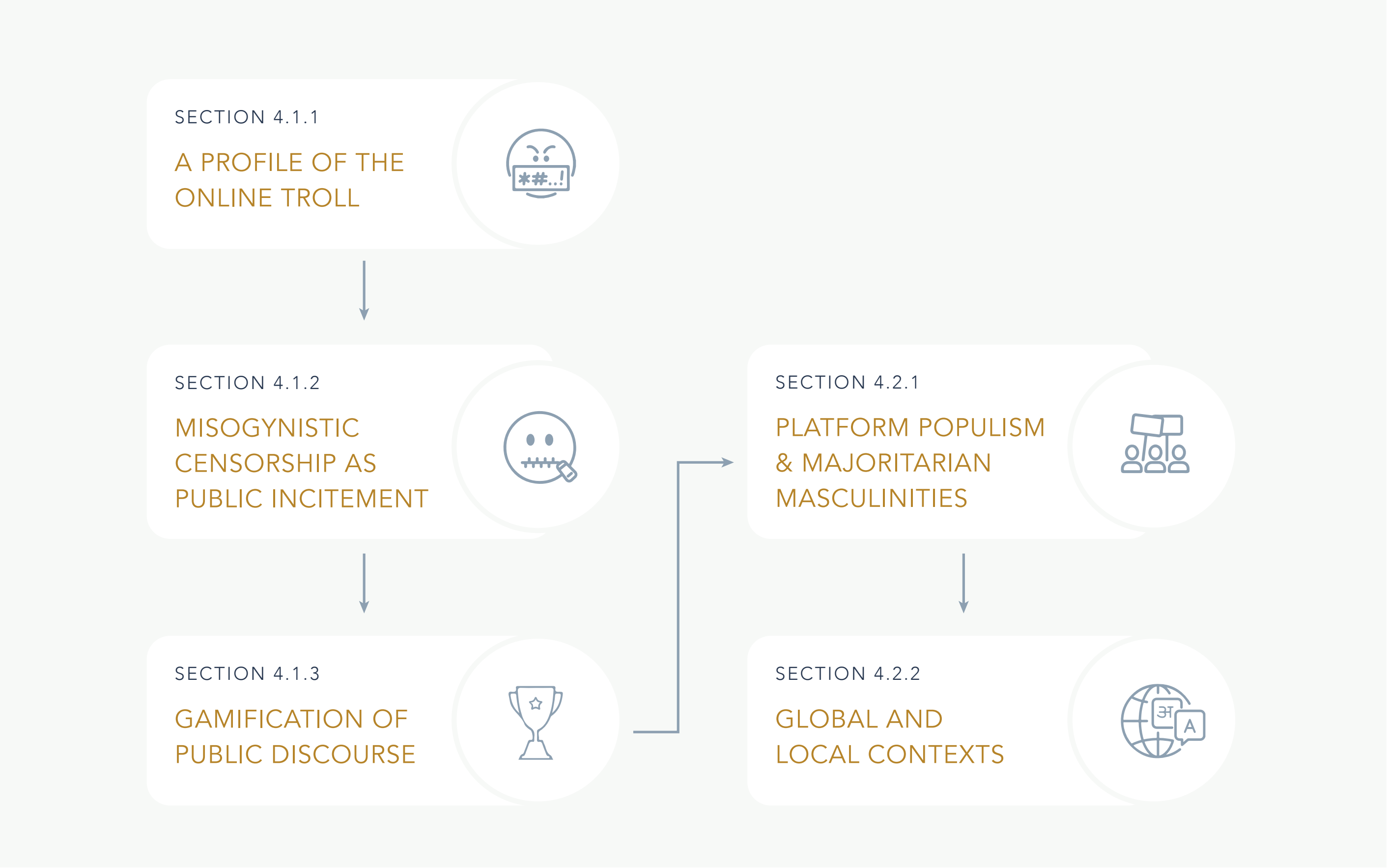
But before we begin to ask these questions, it is crucial to understand the figure of the online troll. Accordingly, section 4.1 begins by unpacking this figure. We rely on Sahana Udupa's conceptualisation of *gaali* (Hindi word for abuse)[36] to show how the peculiarities of the Indian context shape forms of expression and abusive speech.

In addition, we refer to Raminder Kaur and William Mazzarella's work on gendered censorship in India[37] to understand how tactics of incitement and provocation are used by trolls to hijack the attention of *users* and gain control of the public narrative surrounding an event.

We then conclude the first part of the section by considering how the affordances of social media platforms (dis)incentivise certain forms of interaction, usually to the detriment of the health of the public conversation.

The second part examines the role of social media platforms as immensely significant and powerful actors in the modern political economy. Building on the insights in the first part, we consider how the operation of these inscrutable privately-controlled corporate spaces that constitute the ground of public communication have left us with a crisis of publicness. We then consider how the platformised public sphere colludes with the divisive forces of majoritarianism and populism to drive a regressive gender politics that encourages misogynistic trolls to engage in gendered and majoritarian violence.

The section concludes with a discussion on the problem of unevenness in, and inadequacies of, content governance efforts across the globe, specifically in the context of regulating online gender-based violence, and points to the need to devote more resources toward contextually-sensitive approaches to content moderation.



SECTION 4.1.1

A PROFILE OF THE ONLINE TROLL

SECTION 4.1.2

MISOGYNISTIC CENSORSHIP AS PUBLIC INCITEMENT

SECTION 4.1.3

GAMIFICATION OF PUBLIC DISCOURSE

SECTION 4.2.1

PLATFORM POPULISM & MAJORITARIAN MASCULINITIES

SECTION 4.2.2

GLOBAL AND LOCAL CONTEXTS

Infographic 6: Analysis of online misogynistic speech and the affordances of social media. ***Click within the graphic to go to the relevant section.***

4.1 Destabilising the concept of (online) violence

Based on our findings and the wide gamut of abusive behaviours 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 destabilise and problematise the concept of online gendered violence.

As detailed in section 3, the forms of violence against women were not limited to the use of profanity or violent invective. Instead, they often took on more subtle forms such as references to the supposed authority of their husbands or fathers whom they had 'dishonoured,' mentions that infantilised women and trivialised their political careers, and insinuations about the intelligence or qualifications of Dalit women, to name a few common tropes.

We take this central insight about the variability of abusive speech practices as a foothold to launch into our analysis.

4.1.1 A profile of the online troll

In this section, we deal with the online troll specifically in the realm of politics. This class of largely anonymous political actors, also referred to as troll armies, keyboard warriors, and "Twitter patriots,"[38] have been identified all over the globe. What these monikers share is a vision of participatory democracy where ordinary citizens can make their voices heard through collective engagement and contribute to political agenda setting.

This rather positive vision has, unfortunately, remained precisely that—a distant utopic vision. Today, the accepted definition of a troll describes a series of negative attributes (see Infographic 7); a troll is "A person who intentionally antagonises others online by posting inflammatory, irrelevant, or offensive comments or other disruptive content."[39]

AN ONLINE TROLL IS

A person who intentionally antagonises others online by posting inflammatory, irrelevant or offensive comments or other disruptive content.

TACTICS:

Exploiting algorithmic affordances of virality and anonymity to amplify hate and disinformation.

Using social media to carry out orchestrated attacks and coordinated harassment campaigns intended to crowd out other voices.

Exhibiting inauthentic user behaviour to give false impressions about mass political engagement.

SPEECH REGISTERS:

Using the device of comedy to thinly veil misogynistic abuse.

Couching abusive messages in seemingly harmless and deeply embedded cultural references.

Posting extreme and provocative messages with a detached nonchalance.

Infographic 7: Distinctive tactics and speech registers of online trolls

Political parties across the globe now widely recognise the importance of social media in their campaigning strategies. The information technology (IT) cell of the ruling Bharatiya Janata Party (BJP) in India is credited with playing a significant part in orchestrating the party's winning strategy in the 2014 elections, using social media to control the narrative.[40]

The strategy brings to mind the practice of "astroturfing," a term coined by David Bandurski to describe the Chinese Communist Party's practice, as far back as 2007, of recruiting college students as "web commentators" to post pro-party messages on forums and message boards in order to give the false and misleading impression of grassroot support for the party.[41] Political astroturfing strategies have since come a long way, with the increased presence of bot accounts on social media, and highly organised and coordinated harassment and disinformation campaigns. Our research found a significant presence of this manner of inauthentic user behaviour that produced pro-government content and engagement; indulged in coordinated attacks against women dissenters, journalists, and members of opposition parties; and amplified the algorithmic reach of such content. The most common tactic we observed was decidedly unsophisticated in its approach, with multiple anonymous accounts simply retweeting one abusive tweet in reply to a tweet by one of our subjects. In effect, the same message was copied and posted multiple times, one below another, by different anonymous accounts. These forms of targeting are not veiled or subtle attempts to evade detection, but rather, a straightforward method of amplifying messages of hate on the platform.

Though trolls are not always affiliated to political parties, they do see social media as a site of radical disintermediation, where their voices can be heard by the powers that be, and where they can enact their political vision. Udupa examines the intersections of emergent practices of online abuse with the broader aspirations of political participation among a younger generation of Indian users. She develops the emic concept of *"gaali"* to capture the interlocking practices of insult, comedy, shame, and abuse that unfold in a blurred arena of online speech. She argues how

> *"on this slippery ground of shifting practices, comedy stops and insult begins, or insult morphs into abuse in mutually generative ways."*[42]

In making the links between online abuse and aspirations of political participation, Udupa offers some crucial insights. One of these is the recognition that online trolling practices lie on a continuum, and that it is not possible to make neat demarcations between protected, offensive, abusive, and dangerous speech.

Another insight is that online abuse has a deeply gendered structuring—the raking up of "the private" and sexual accusations represents a re-politicisation of the "domestic sphere" through a masculinist logic of shame with an intention of intimidation.[43] These insights are affirmed by our findings, some of which are covered in section 3.1. An overwhelming amount of trolling directed at the women we studied made some mention of their private lives or their family, or about how they had brought shame to their husbands or fathers. It was also clear from the tone of many of the messages, that users considered abusive and misogynistic posts on social media as the only way that they could make their political voices be heard by the powers that be.

Our research also showed the ways in which trolls used *gaali* and the device of comedy to veil abuse. This mode of address used by trolls, which Whitney Phillips and Ryan Miler have called the "aspirational register of lulz,"[44] refers to a way of engaging with discourse in which nothing is to be taken too seriously, and anything is fair game to be ridiculed. Phillips and Milner write that this "chorus of ironic, nihilistic, fetishistic laughter [has] created the perfect conditions for bigotry to spread stealthily [...] like harmless fun."[45] Most of the trolling that we observed was also of this nature—animated by a particular kind of collective misogynistic laughter that poked fun at women and celebrated their distress.

4.1.2 Misogynistic censorship as public incitement

Problematising the idea of online violence against women is important not only to destabilise ossified notions of 'obscenity,' 'hate speech,' and 'incitement to violence' in the law, but also to unravel the more amorphous set of rules or social scripts around which practices of cultural regulation that seek to police women take shape.

One entry point into understanding this is through ideas of censorship. India's recent past is replete with examples of films and literature that have been banned or censored, citing the rationale of protecting women.[46] Kaur and Mazzarella incisively argue that something deeper is at play in the publicised spectacle of banning a book or prohibiting the release of a film. They suggest that there does not exist a direct or straightforward relation between the act of the censor and the desired outcome of suppressing the censored article. The public drama and moral outrage around the censorship of a sex scene in a film, for example, foregrounds rather than conceals the prohibitive act of censorship, turning it into a public spectacle. Depending on one's perspective, it becomes either a story of depraved youth or of prudish intolerance. Put differently, the act of censorship does much more than just suppress the 'objectionable' content; it drives public conversation and discourse around the politics of obscenity, decency, propriety, and respectability in society. As Kaur and Mazzarella write, "censorship is not in but of the public sphere."[47]

To elucidate this, we could consider an example on social media. A particular scene from the 2018 Hindi film *Veere Di Wedding* became the subject of controversy and public debate, offline and online. The scene in question depicts actor Swara Bhasker's character masturbating. This was seen by certain conservative audiences as an affront, and an unacceptable portrayal of womanhood in Indian cinema. In the years since the release of the film, Bhasker, an outspoken woman who is highly engaged in current political debates and active on Twitter, continues to be incessantly trolled for this supposed transgression. Bhasker has said of the trolling,

> *"I can't even post a photo of a flower without people linking it to masturbation or referencing ungli (finger) after Veere Di Wedding came out. It's ugly and amounts to cyber sexual harassment but I feel very strongly about not succumbing to online bullying or limiting my presence online because of it."*[48]

The policing or silencing underlying the censorship of a supposedly unspeakable moral transgression by a woman—regarded in the law as "lascivious" and "appealing to prurient interest"—is an important dimension of the story, but not the entire story. The other dimension has to do with what censorship generates or activates, rather than what it silences or suppresses. In the context of social media, it also has to do with how the particular configurations of medium and message create and reproduce public discourse. As a form of cultural and gendered regulation, pervasive online misogynistic speech and trolling instrumentalise the algorithmic affordances of social media platforms to hijack the public conversation, and forcefully evacuate narratives that unsettle Brahminical patriarchy through tactics of shaming, provocation, and incitement.

These discursive practices of censorship hence routinise a pattern of incitement in the online public sphere. These practices tend, by and large, to restrain the sexual and political autonomy of women, something that our findings on patterns of abuse detailed in Section 3.1, also attest to.

4.1.3 Gamification of public discourse

Central to the formation of the figure of the online troll is the techno-design of social media platforms. The gamification of public discourse online is rooted in a moral vision that is narrow and limiting. On Twitter, for instance, likes, retweets, impressions, and other quantitative metrics of engagement become the principal factor determining the spread and success of a message.

C. Thi Nguyen notes how the gamification of public discourse serves a variety of purposes.[49] A game designer, he writes, not only makes the rules of the game, but more importantly, also determines how players act within the game, and what goals they strive toward. This is perfectly harmless in the context of achieving a mission in a game, but when transposed to the realm of public discourse, it has dangerous real-life consequences. Gamification of public discourse instrumentalises social and moral values, placing value blinders that restrict our field of vision (through the operation of echo chambers and filter bubbles), and revealing only that which enables us to achieve our narrow goals. By reducing the field of politics to something which is simple and knowable, the gamification of public discourse negates complexity in online debates and conveniently masks "the hideous existential nuance of the world."[50]

Thus, the design mechanics of social media platforms play a defining role in the way online political discourse plays out. They are engineered in specific ways to encourage certain kinds of interaction. The choreographed drama characterising the online public sphere presents women with an impossible tradeoff: they have to either be prepared to keep receiving abuse or reduce their communicative autonomy by making their accounts private, hiding comments, disabling retweets, and blocking accounts. Shehla Rashid, a human rights activist, captures this double bind:

> *"As women we have come to terms with the fact that our freedom of speech is qualified. If we have fame or reach, there will be a flipside to this. We will face abuse on social media."*[51]

It is important to recognise the ways in which the infrastructural affordances of social media platform, such as anonymity,[52] likes, and retweets, actively determine how users on the site interact with one another. Michael Walsh and Stephanie Baker's research on user strategies of dealing with hostility on Twitter has shown that in dealing with hostile communication situations, a majority of users choose the path of avoidance as a way of "saving face" on the platform, rather than engaging in restorative interactions.[53] These strategies involve the use of pseudonyms, multiple accounts, and other ways of minimising exposure such as protecting their tweets and limiting their reach.[54]

The findings from our study corroborate this conclusion about avoidance being the most common strategy. Despite the prominence in public-political life of the women we studied, or perhaps because of it, we found no evidence of them either engaging with abusive mentions in their replies or using restorative interactions in the form of counter-speech. One can only surmise that the choice of avoidance as a strategy is just as much a matter of saving "Twitter-face"[55] or public impression management, as it is of maintaining one's sanity in the face of a deluge of hateful messages.

Despite public admissions by platforms such as Twitter that they need to "do better"[56] about women's safety, it is evident that hate speech and trolling at scale is a feature and not a bug of platform design. Techno-design features enable trolls to employ violent tactics that have a deeply gendered structuring such as harassment, doxxing, pile-ons, and ratioing.[57] Aggressors are afforded the publicity they desire, while their anonymity shields them from facing any consequences. The targets of the aggression, on the other hand, are often backed into a corner and rendered hyper-visible.

In conclusion, destabilising the notion of censorship (as something more than simply the prohibitive act) allows us to reconcile the simultaneous operation of seemingly incommensurate forms of cultural regulation—publicity and censorship. Our findings show how these discursive practices coalesce in specific ways on social media to routinise a pattern of incitement and reward bad faith actors.

Our research also shows how the constant trolling and violence directed at women in the online public sphere tend to be highly choreographed affairs that shape the rhythms of public debate online, following the predictable peaks and troughs of public outcry and moral indignation. In this familiar script, it is invariably women who are assigned the role of docile subjects of patriarchal control, and any transgression of these rigid norms is met with trolling, abuse, and conservative moral policing.

4.2 Platform as structure

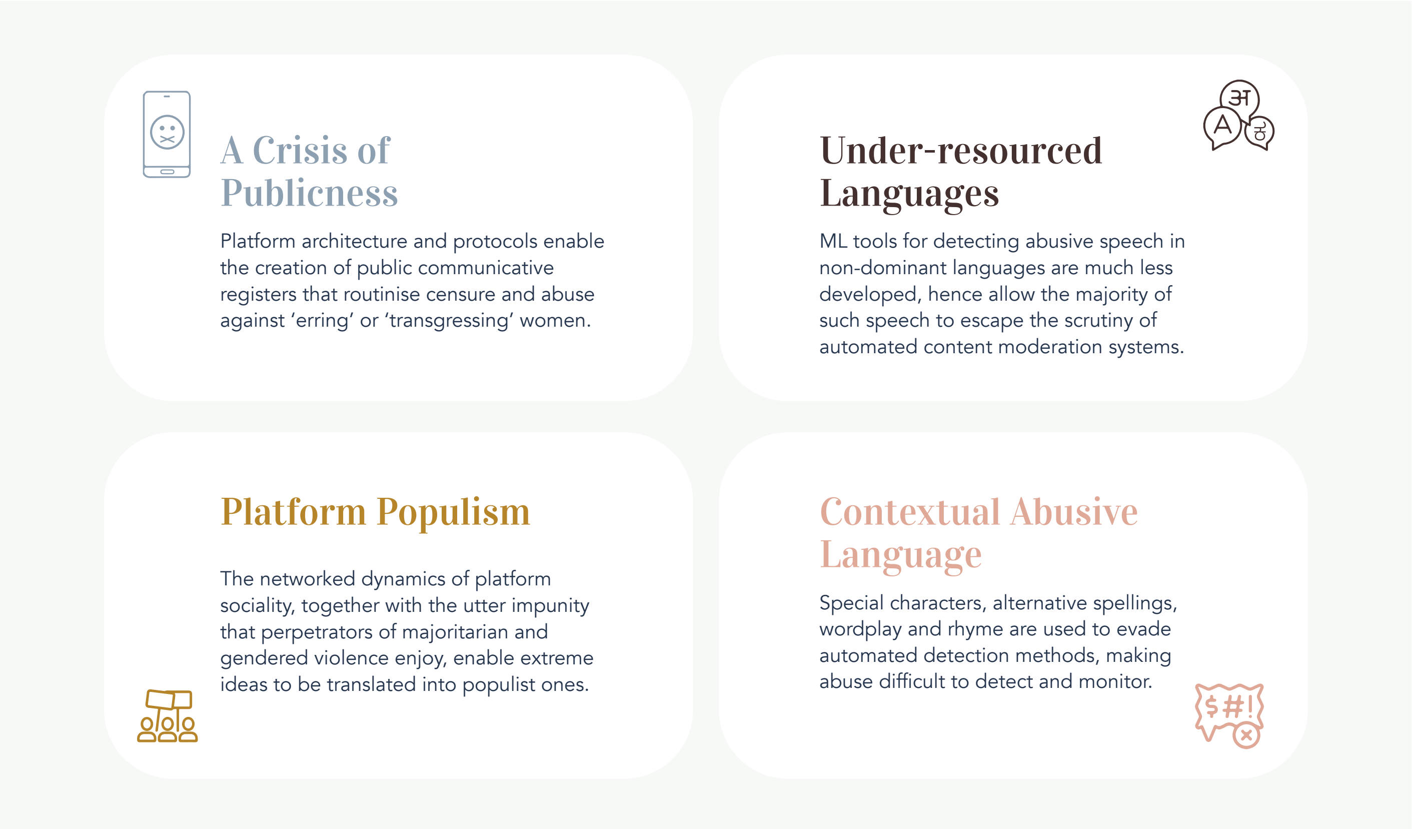
In the context of the rise of majoritarian politics in India and unrestrained abuse directed at the women in our study, this section will explore the linkages between powerful and privately-controlled social media platforms, the abuse received by women politicians online, and the Indian state. Central to this account is an examination of the modern public sphere and the ways in which platformised social environments facilitate a perverse publicity, marking, targeting, and making an example out of women who challenge patriarchal norms.

In this analysis, we have been interested in a 'publicness' that is more expansive[58] than the traditional Habermasian concept of communicative rationality or of the public sphere as inhabited by the citizen as rational-critical political actor. Rather than such a restrictive and formalist theory of public action and communicative activity, we are interested in a publicness that references other emergent forms of public conduct (here, trolling) that constitute political action in a different register. Publicness, writes Slavko Splichal, refers to

> *"a specific mode of relationship among people based on visibility and access, which is essential for the processes of collective self-understanding and constitutive to democratic societies."*[59]

The right to publicness, then, is independent of the formal categories of politics, the State, or legal frameworks that secure participatory rights. It is a fundamental, pre-institutional aspect of public action, and a precursor to the right to communicate. In the context of the silencing effect of online trolling, the right to publicness, as opposed to most elaborations of the legal right to freedom of expression or public participation, would emphasise that the right to be heard be placed on an equal footing as the right to speak.[60] In our analysis of hateful, abusive, and problematic speech online, we have explored how this publicness has been subverted and instrumentalised by platforms to drive a regressive gender politics. While the policing of women in public spaces is intrinsic to patriarchal sociality, corporatised platform environments add a new dimension by providing a means to weaponise and instrumentalise affect.[61] Platform architecture and protocols enable the creation of public communicative registers that routinise censure and abuse against 'erring' or 'transgressing' women. The figure of the troll and the mass spectacle of trolling seek to set the terms of communicative discourse, instrumentalising (network) affinity and virality for a gendered restructuring of public space.

The publicness of the public sphere, both in terms of what occurs in open view and what may concern the multiplicity of communities comprising the public, is thus hollowed out, as inscrutable corporate-controlled spaces and rules of interaction entrench a new modus operandi of everyday discourse. Social media platforms, thus, need to be conceived not just as the new infrastructures of communication, but as structures that fundamentally reorder social relations, communicative protocols, and the very terrain of politics. With this conception of publicness as the normative basis of the right to communicate, we examine the rising tide of online majoritarian and gendered violence on social media platforms and their interconnections. Infographic 8 offers a brief summary of the structuring role that social media platforms play in modern-day political formations.



A Crisis of Publicness

Platform architecture and protocols enable the creation of public communicative registers that routinise censure and abuse against 'erring' or 'transgressing' women.

Under-resourced Languages

ML tools for detecting abusive speech in non-dominant languages are much less developed, hence allow the majority of such speech to escape the scrutiny of automated content moderation systems.

Platform Populism

The networked dynamics of platform sociality, together with the utter impunity that perpetrators of majoritarian and gendered violence enjoy, enable extreme ideas to be translated into populist ones.

Contextual Abusive Language

Special characters, alternative spellings, wordplay and rhyme are used to evade automated detection methods, making abuse difficult to detect and monitor.

Infographic 8: Platform as structure

4.2.1 Platform populism and majoritarian masculinities

This section situates our findings within the growing body of scholarship on the forms of mediated populism in postcolonial democracies and the complex ways in which networked citizens negotiate dominant political imaginaries. We begin by moving outward from the analysis of platform affordances, which structure online conduct, and taking a broader view of the role of the platform within larger socio-political formations. We discuss the tactical collusion between populism and platform sociality; the legitimacy afforded by platforms to acts of public, majoritarian violence; and the ways in which acts of gendered violence are connected with the promise of inclusion into an empowered majority, with immunities from prosecution.

On the ubiquity and pervasiveness of majoritarian violence in contemporary India, Thomas Blom Hansen writes that it is this "sense of freedom when in a crowd, or a sense of having been given 'permission' by one's leaders to act, to hit, and to abuse that are the most powerful ingredients in public violence today."[62] Our research shows that political dissenters, Muslim women, and those from opposition parties were at the receiving end of a disproportionate amount of abuse on Twitter. Social media platforms, therefore, occupy a central role in the reconfigured field of modern politics, and are, in part, answerable for the growing cloud of communal disharmony, religious intolerance, and the silencing of dissent in India. Central to our perspective is the recognition that social media platforms do much more than passively mediate online political engagement and conduct.[63] Rather than occupying a passive or relaying role, social media platforms play a determinative and generative role in the domain of politics, and engender fundamental shifts in modern ways of "doing politics."[64]

The rampant spread of online misogyny is therefore not a secondary or residual phenomenon, but a structural characteristic of modern-day political formations. Put differently, the concurrent rise of right-wing populism and the growth of online toxic-masculinist technocultures are not coincidental, but rather, colinear developments that work in tandem to constitute new political subjectivities.[65] We observed numerous instances in which the rhetoric of misogyny and right-wing ideologies coincide, and the ways in which the adjacent discourses of gender and nation are both affected and enabled by the platformised online public sphere—a media landscape characterised by speed, virality, and an unending stream of novel content competing for attention.

This point about the unprecedented distribution logics of platform sociality assumes central importance in the final section (section 5), where we consider how the law can respond to the problem of online misogynistic speech. The networked dynamics of platform sociality, together with the utter impunity that the perpetrators of majoritarian and gendered violence enjoy, enable radical ideas to be translated into populist ones, shifting the Overton window[66] of what is considered acceptable speech in public.[67] Indeed, the normalisation of misogynistic speech online has been a central component of the right-wing political playbook in India.[68] The accommodating attitude that the country's elected representatives have toward the daily barrage of online violence directed at their political opponents effectively constitutes an endorsement of such abusive behaviour.[69] This toxic culture of muscular nationalism and misogyny pervades online spaces and emboldens misogynistic trolls to use any means at their disposal in pursuit of their political ends. For instance, we encountered numerous violent mentions attacking women in our sample that simultaneously directed jingoistic praise at the Prime Minister.

We found that the term *presstitute* has become a choice insult used by trolls to attack women journalists. Leaked documents have also shown the kinds of instructions and templated tweets that circulate to create the impression of a groundswell of support for the current dispensation's policies.[70] The kinds of left-field tactics that organised trolls use to spread disinformation and hate point to the sheer brazenness and impunity with which they pursue their political ends.[71]

The empirical fact that not all trolls are part of organised collectives or affiliated with political parties is not relevant to the larger point we make here. That is, the visible protections and immunities afforded to those who hurl abuse at women as ways of exercising their political voice, is all the license, invitation even, needed to indulge in such behaviour. In this context, the continued silence of the incumbent political authorities on the issue of online misogyny is deafening.[72]

Furthering the argument in section 4.1 about the infrastructural affordances of the platformised public sphere and the kinds of online conduct they incentivise, we have discussed in this section how the public performance of gendered violence online has become a crucial component of contemporary political formations. By inducing this kind of political engagement, social media platforms become a medium in which the conduct of contemporary politics shifts towards a hegemonic populism rooted in regressive and misogynistic values. This fundamental restructuring of the terrain of the political has normalised and banalised the discriminatory treatment of women in the online public sphere.

4.2.2 Global and local contexts

As we have observed through the course of this report, the nature of violence directed at Indian women on social media platforms is highly contextual and culturally specific. The distinct ways in which cultural difference is negotiated within the logics of globalisation and platformisation in emerging regional markets is a vital axis for any inquiry mapping social media; all the more, where it pertains to questions of social power. Any study of hateful or abusive speech also needs to account for the regulation of harmful content in local languages.

Unlike languages like English or French, which have highly sophisticated natural language processing (NLP) algorithms developed for them, speakers of non-dominant and under-resourced languages are systemically excluded from such governance modalities and disproportionately subjected to algorithmic harms. Mashinka Hakopian refers to this phenomenon as "algolinguicism"—

> *"A matrix of automated processes that minoritise language-users outside the Global North and obstruct their access to political participation."*[73]

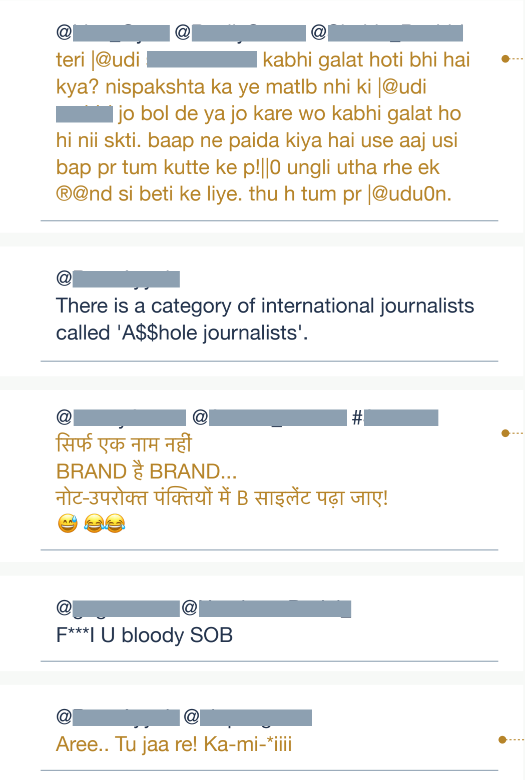
These language gaps in content moderation systems allow abusive posts to spread unchecked and have had cataclysmic effects on countries across the globe.[74]

The lack of cultural specificity in content moderation is exemplified by platforms' terms of service, such as Facebook's Community Guidelines and the Twitter Rules, which are universal in their scope. An overarching feature of these terms of service is that they borrow from (usually American) legal frames of legitimacy, most notably the commitment to free speech, as justifications for their universal scope.

While they do, in principle, contain positive rules regarding the intolerance of hateful and abusive content, in practice, they are woefully ineffective in actually enforcing these standards in accordance with local speech contexts. These terms of service, therefore, end up reading as little more than moral platitudes and parchment barriers. More specifically, a part of the problem of online gender-based violence that sets it apart from other policy issues relates to the task of detection of such content.

Our research shows that trolls employ different tactics such as the use of special characters, alternative spellings, and even the use of wordplay and rhyme to evade automated detection methods such as keyword filtering. This makes abuse, harassment, and trolling of the nature that we have observed as part of this study especially difficult to detect and monitor.

SNAPSHOTS



@[redacted] @[redacted] @[redacted]
teri |@udi [redacted] kabhi galat hoti bhi hai kya? nispakshta ka ye matlb nhi ki |@udi [redacted] jo bol de ya jo kare wo kabhi galat ho hi nii skti. baap ne paida kiya hai use aaj usi bap pr tum kutte ke p!||0 ungli utha rhe ek ®@nd si beti ke liye. thu h tum pr |@udu0n.

(Hindi) Is this bitch ever wrong about anything? Non-partisanship does not mean that whatever she says or does can never be wrong. You sons of bitches are pointing fingers at the father instead of this whore of a daughter. Shame on you dicks.

@[redacted]
There is a category of international journalists called 'A$$hole journalists'.

@[redacted] @[redacted] #[redacted]
सिर्फ एक नाम नहीं
BRAND है BRAND...
नोट-उपरोक्त पंक्तियों में B साइलेंट पढ़ा जाए!
😅 😂😂

(Hindi) REDACTED is not just a name, but a BRAND.. Note: the B above is to be read silently.
["Rand"(रांड) means prostitute in Hindi.]

@[redacted] @[redacted]
F***I U bloody SOB

@[redacted] @[redacted]
Aree.. Tu jaa re! Ka-mi-*iiii

(Hindi) Aye, get out, *kameeni.*
[The word "kameeni" is a casteist slur.]

Direct and unveiled terms like *randi* (whore) or '*prestitute*,' which are widely used exclusively to attack and demean women, continue to be used with impunity on the platform. We also observed the unabashed use of casteist slurs, such as *chamar* and *bhangi*, that have been officially recognised and proscribed under Indian law.

We do not argue here for a simplistic ban of these offensive terms on the platforms. As Tarleton Gillespie points out,[75] this would be a misguided and myopic approach because it would ignore people using slurs in a reclaimed fashion for an "in-group" they are part of; activists documenting groups that hold those positions to 'name and shame'; and people who devise elaborate vocabularies to talk about them in code or couched language.

Instead, the unabashed use of such offensive terms is indicative of the need to develop more contextual content moderation practices that can use machine learning tools to augment the capabilities of human moderators equipped with knowledge of local events and the cultural competence to be able to make these fine distinctions. Human moderators are not only vital, but their sensitivity to changing cultural mores is a prerequisite, given that hateful speech often takes the form of dog whistles that couch prejudice in seemingly innocuous phrases. For example, in our research, we observed the use of the term "ola uber," a corruption of *Allahu Akbar* (a common phrase used by Muslims in various situations, including in prayer), as code to refer to Muslims.

One major problem relates to the design and construction of Twitter's rules, which are universalistic in scope. This approach comes at the expense of local and contextual governance approaches. Robyn Caplan writes that there are broadly three content moderation strategies adopted by platforms of different sizes, each of which balances the interests of context and consistency in the application of their rules differently.[76] These three strategies are artisanal, community-led, and industrial approaches.

Typically, the artisanal approach is used by smaller platforms and involves case-by-case governance by a smaller team of moderators. As the name suggests, community-reliant approaches involve the active contribution of volunteer moderators, in combination with platform policy. The industrial approach to content moderation is employed by big platforms with larger user bases. The work of moderation in these companies is performed by large moderation teams of tens of thousands of workers who are employed to enforce rules made by a separate policy team.

With their massive user bases across the globe, large social media companies such as Facebook, YouTube, and Twitter employ the industrial approach to screen massive amounts of content on their platforms. While each of the strategies have their own strengths and weaknesses, we highlight below just how inadequate the current industrial moderation model used by large social media platforms is in regulating online gender-based violence.

Inferences from the data in the compliance reports of significant social media intermediaries (SSMIs) in India for March 2022[77] show that Facebook India falls short in its proactive monitoring solutions for bullying and harassment. Partly due to the reasons outlined above, the technical state-of-the-art of machine learning technology is not equipped to effectively detect content in this particular policy area. Facebook India's proactive detection rate for the bullying and harassment category is 79.1%, which is low in comparison to all other categories, except hate speech, which, for similar reasons, also has a low detection rate.

Similarly, Twitter India's compliance report for the same reporting period shows that the platform receives the most number of user grievances for abuse and harassment, although it does not provide granular data about its proactive monitoring efforts for this category. Extrapolating from these disclosures made by these two different platforms, we may infer that a Venn diagram of the policy areas where proactive detection tools are ineffective and where platforms receive the most user complaints is a near-perfect circle.

This laxity also extends to the enforcement of platform rules. The case of journalist Kavin Malar exemplifies the lack of seriousness with which social media platforms enforce their rules in cases of harassment of women. In August 2020, Malar reported a user for posting two photographs of her on Facebook with an offensive caption: "My rate is 1000 rupees."[78] Soon after the post was uploaded, she began to receive calls and obscene messages from other users harassing her. In response to Malar's complaint, Facebook refused to take action against the user, stating that the post was not against their community standards. It was only after Malar posted a photo of the platform's response on her Twitter account, leading to a public outcry, that Facebook relented and suspended the user's account.

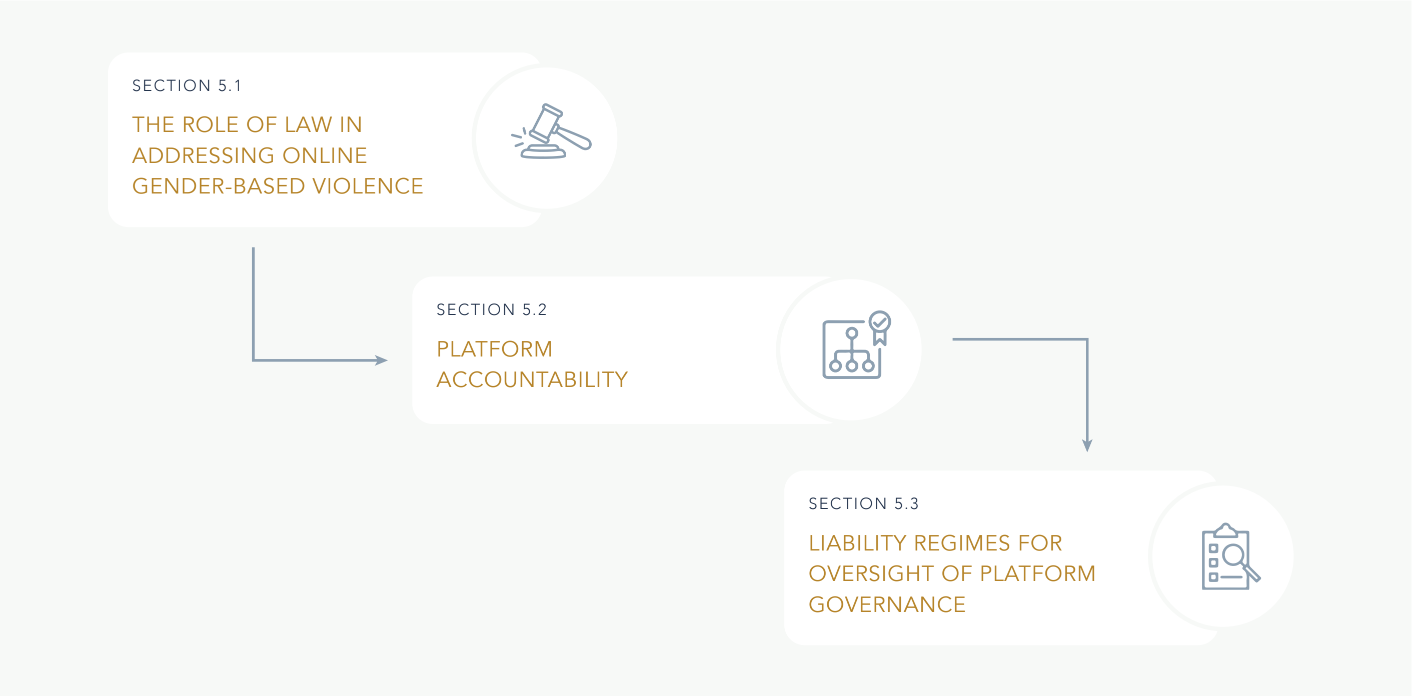
Platforms as large as Facebook or Twitter, certainly need to make tradeoffs in choosing a moderation strategy that can balance the interests of context and consistency. However, the evidence from their compliance reports in India reveals that when it comes to gender-based hate and trolling, the odds are overwhelmingly stacked against women, both in terms of detection of harmful content as well as the enforcement of platform rules in relation to such content.

All of the above point to the need to devote more resources toward contextually-sensitive approaches to content moderation. This can both enhance the rate of detection—for instance, by training local human moderators and developing machine learning tools—as well as improve reporting mechanisms and responsiveness to user grievances. User complaints must be utilised 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.

05

Legal-Institutional Responses

This concluding section returns to some of the central questions that animate the research. We start by exploring the arguments for and against the use of the law to combat gender-based violence before moving on to discuss how the right to free speech is to be understood in the context of the platformised public sphere. The section concludes by considering possible institutional arrangements for effective platform oversight, making recommendations for transparency reporting, grievance redressal, and the management of amplification/virality.



SECTION 5.1

THE ROLE OF LAW IN ADDRESSING ONLINE GENDER-BASED VIOLENCE

SECTION 5.2

PLATFORM ACCOUNTABILITY

SECTION 5.3

LIABILITY REGIMES FOR OVERSIGHT OF PLATFORM GOVERNANCE

Infographic 9: Legal-Institutional responses. ***Click within the graphic to go to the relevant section.***

5.1 The role of law in addressing online gender-based violence

Given that online gender-based violence covers a broad spectrum of behaviours, ranging from acts of trolling to more egregious offences such as cyber stalking, doxxing or non-consensual intimate image distribution, we urgently call for a broader constitutional vocabulary to describe gendered hate speech. This cannot be based on patriarchal notions of modesty, decency or public morality, and instead, must be premised on the recognition that such pervasive sexist speech reinforces structures of oppression and discrimination along the lines of gender.[79] In keeping with the primary focus of this study, we limit our analysis to cases of trolling, abuse, and misogyny of the kinds we have reported on in the preceding sections.

The Me Too movement, founded by American activist Tarana Burke in 2006, exploded into the global mainstream when the hashtag #MeToo began to trend on social media. The global feminist movement which took the world by storm in 2017, was driven in large part by the amplification powers of social media, in addition to a significant offline engagement. Even as it was celebrated as a cultural moment of reckoning, there were issues where a feminist consensus was difficult to achieve. At the heart of the disagreement was the definition of sexual harm which could include a spectrum of behaviours, ranging from serious forms of sexual misconduct to cases that were more difficult to judge, often involving microaggressions or subtle forms of gendered violence.[80]

Brenda Cossman writes that we can read the feminist contestations around #MeToo and the governance of sexual speech as the intellectual inheritance of the sex wars of the 1970s and '80s that centered around the issue of pornography.[81] She shows that there are important resonances in the political positions of the anti-porn and #MeToo feminists versus the pro-sex feminists and feminist detractors of #MeToo, respectively. In both movements, the feminist disagreements around sexuality, consent, and the definition of sexual harm are centered around the axes of pleasure/danger and agency/victimhood. Of course, it is not strictly a question of either/or but rather one of emphasis. #MeToo feminists emphasise the need to recognise the pervasiveness of harmful and insidious sexual behaviours, while feminist detractors of #MeToo raise concerns about the movement's failure to see sexuality as a site of women's pleasure and autonomy. These feminist debates capture the fine contours of the discourse on the role of the law in addressing online gender-based violence.

As the body of knowledge that is held to have the sole power to define and adjudicate sexual harm, the institution of law casts a long shadow over discourses of sexuality and its regulation. In the context of the #MeToo movement, for instance, the slogan "believe all women" and questions of due process became central to disagreements precisely because of the law's discursive power to sit in judgement over the truth claims of sexual harm. In the Indian context, these debates over the role of law came to a head prominently in the feminist disagreements around the "List of Sexual Harassers in Academia," or LoSHA, published on Facebook.[82] According to Cossman,

> *"Loosening law's hold on the definition of sexual harms could go some way to allowing these deeper feminist conversations, in ways that both allow for an affirmation of sexual harm, without endorsing a carceral state."*[83]

This reparative stance[84] that Cossman suggests in the context of the #MeToo movement, also helps to take seriously and reconcile two supposedly incompatible feminist positions underlying this report.

The first is the claim that the ubiquity and normalisation of trolling and misogyny in the online public sphere is a matter of grave public importance, and that the law has an important role to play in addressing these online harms. To build on the assertion made earlier in this report about the need to destabilise the concept of online violence, we rely on speech act theory to conceptualise the harms of pervasive online gender-based violence as relevant to law. This scholarship argues that it is not enough to look only at the content of speech or its effects, but also at the actions constituted by it. It also underscores that the legal recognition of online gender-based violence would constitute an authoritative speech act to counter the pervasiveness of online misogynistic speech.

The second feminist position that is relevant here, is the claim that while social media may be a site of danger where women receive threats and abuse, it is equally a site of pleasure and fulfilment where women can exercise their autonomy and freedom of expression. Thus, paternalistic and overbroad legislations enacted to better women's situation online could, and often do, have the counterproductive effect of further restricting their agency by making undue incursions, and policing or moralising forms of expression.

The friction between these two positions exhorts us to arrive at a set of proposals and recommendations regarding online gender-based violence that target platforms rather than individual offenders. As has been discussed, platforms play a determinative role in shaping public discourse by (dis)incentivising certain kinds of behaviours. Therefore, they need to be bound by rules that recognise the immensity of this responsibility. Building on this submission, the subsequent sections argue the perils associated with legal efforts to police online gender-based violence (specifically, trolling and other speech-based forms of violence) based only on its locutionary content. Instead, our recommendations for regulating user-generated speech are based on the tenet of platform accountability, with increased liability on platforms for online harms. We also discuss some considerations that need to be taken into account in framing such a legal approach and possible regulatory models.

5.1.1 Conceptualising the harms of online misogyny

Acts of trolling reported in this study cannot be construed as milder forms of offensive speech, deserving of legal protection. Indeed, choosing to regulate only the most virulent forms of gendered violence and stepping back from the issue of trolling would result in overlooking one of the most pervasive and insidious forms of gender discrimination online. Pervasive misogynistic trolling online is evidence of how the prevailing legal position has normalised the silencing of women's voices in the digitally-mediated public sphere.

However, saying that forms of trolling that derive their potency from their volume and frequency are not deserving of the status of legally-protected speech is quite different from saying that such speech should be criminally actionable. Criminal law (quite rightly) requires a much higher evidentiary threshold in order to establish harm, and these are unlikely to be met in cases of trolling. The standard of grave and imminent danger, or of incitement is not suitable to assess cases of trolling. How then might we conceptualise the harms of online gender-based trolling, not only as causing individual offence or wounding sentiments, but in ways that are relevant to the law and attentive to the power of such speech to legitimise the discriminatory treatment of women? Austinian speech act theory provides a way of approaching this question.[85] This lens was used most influentially by feminist legal scholar Catherine MacKinnon to open up new ways of thinking about the relation between law and feminism.

By dismantling the speech/action dualism, speech act theory seeks to investigate the different communicative dimensions at play in a linguistic utterance. The speech act is a unity of three different dimensions—the locutionary act (conveyance of the literal meaning), the perlocutionary act (the effect of the utterance on the listener), and the illocutionary act (the action constituted in saying something).[86] J.L. Austin argues that there is a tendency to consider the content of a linguistic utterance and its effect on hearers, but overlook the illocutionary aspects of speech, or the actions constituted by speech. We are chiefly concerned here with the illocutionary dimension of speech, that is, what actions are constituted in certain forms of speech? The paradigmatic example used by Austin to convey this power of a speech act to 'do' things is the performative utterance 'I do' in the context of a wedding ceremony, where the phrase does more than indicate the willingness of the bride and groom; the illocutionary act of saying 'I do' under these conditions constitutes the act of marrying.

The function of speech then is not limited to its semantic content or its effects on hearers. A speech act can carry the illocutionary force of constituting, fortifying or reinforcing social hierarchies. MacKinnon and Andrea Dworkin apply this sort of analysis to demonstrate the broader societal harms of violent pornographic speech which depicts the subordination of women. Their claim is that such porn does not just depict the degrading treatment and subordination of women, it has the illocutionary force of subordinating and silencing women, and is hence constitutive of harm. A critical piece of this feminist political analysis lies in recognising patterns of injustice and establishing connections between how violent porn depicts the degrading treatment of women, and the fact that women are overwhelmingly the victims of sexual violence.[87] Without venturing into discussions of porn and the highly fraught connections between these statements, for the purposes of this study, we use a similar approach to conceptualise how acts of trolling cause and constitute the systemic silencing and subordination of women online.

That women are overwhelmingly the targets of online abuse and trolling is not only an empirical fact, but also a political one. The ubiquity of trolling testifies to the authority that such misogynistic beliefs hold. This may not be taken to mean that the beliefs themselves are held in high esteem, or even that they are espoused by significant political figures (although these do constitute conditions of felicity). Rather, as Arti Raghavan writes, for the hearers that count (online trolls), the authoritativeness of such messages is

> *"derived from the fact that there are countless similar messages on the internet, the cumulative effect of which is to lend authoritativeness to the message, and result in a subordinating effect."*[88]

What is distinctive about social media misogyny is that such views hold authority because the infrastructural context of the social media domain allows them to rise to prominence, and each subsequent abusive message enacts a permissibility fact that implicitly reinforces the rules of acceptable online conduct. While the continuities between offline socio-cultural relationalities and online experiences of abuse are widely recognised, what is often overlooked in feminist scholarship is that the primary context that distinguishes social media misogyny is the illocutionary aspect of speech mobilised through platform power—the regimes of permissibility that capitalist social media platforms create. This does not mean that the embedded socio-cultural antecedents of misogynistic speech acts don't matter, or that they disappear. Rather, they become incidental to the process through which a new normativity of gendered speech and conduct is produced by capitalist platforms that instrumentalise local culture for profit.

Further, as Anjalee de Silva argues, the law's continued silence on the issue of pervasive online gender-based violence is an act of accommodation from which trolls derive authority.[89] The law, as authoritative speech with the institutional backing of the state, has an important affirmative role to play in challenging and countering the permissibility facts of online misogyny.[90] There is tremendous power in attending to the illocutionary force of speech and, in so doing, identifying and naming the structural abuse faced by women online as relevant to law.

As we shall see in the next section, however, our faith in the law has to be tempered because this account of the illocutionary force of a speech act does not always translate unproblematically when displaced into the domain of lawmaking.

5.1.2 Unintended consequences

The radical feminist critiques of pornography, and representations of women more generally, provide an invaluable grid to make legible widespread forms of normalised gendered violence. The emphasis on the use of the law as an instrument to police such violence, however, results in a carceral politics with some troubling consequences. The push toward broader definitions of sexual harm has had the unintended consequences of reinforcing hegemonic patriarchal and heteronormative societal norms. In recent years, with the increase of online gendered violence, we have seen these consequences play out in laws regulating sexual speech, ostensibly intended to protect women in the online sphere.

Consider the recent proposed amendment to the Kerala Police Act which sought to impose up to five years' imprisonment, or a fine of up to Rs. 10,000, for defamatory media posts that were "threatening, abusive, humiliating or defamatory."[91] Initially triggered by an incident where a group of women activists confronted a blogger who used his YouTube channel to broadcast derogatory comments about them, the Kerala government claimed that the amendment would help prevent hate speech and cyber attacks against women and children.[92] Amid public criticism about the vague wording of the provision and warnings that it would have a chilling effect on the right to freedom of speech, the proposed amendment was quickly withdrawn by the state government.[93]

An example from the UK may be especially instructive for our purposes, given its connections to feminist debates of the 70s and 80s which, as discussed earlier, centred around the issue of pornography. In 2014, the British government passed a law that prohibited certain sex acts from featuring in porn (including on online porn platforms) produced in the country.[94] This list includes acts such as spanking, caning, whipping, and physical restraint. As Amia Srinivasan points out, these might be assumed to involve women's subordination, but are in fact, also characteristic of femdom porn. Similarly, facesitting and female ejaculation, which are emblematic of women's pleasure, are also on the list of outlawed sex acts. Srinivasan writes,

> *"What is officially sanctioned here, by virtue of being left off the list, is the most mainstream porn, the porn that turns most people on. But the whole point of the feminist critiques of porn was to disrupt the logic of the mainstream: to suggest that what turns most people on is not thereby OK. To prohibit only what is marginal in sex is to reinforce the hegemony of mainstream sexuality: to reinforce mainstream misogyny."*[95]

These examples show how laws that proscribe certain forms of speech are liable to misuse and tend to reinforce the marginality of vulnerable groups by reinscribing gender and sexual hierarchies. Further, the mere accumulation of rights alone is not a guarantor of progress, especially on the subject of gender justice. In addition to the misuse and misapplication[96] of laws intended to protect women, there remains a gap between the availability of a right and access to a remedy. The legal system is deeply unwelcoming to women who choose to enforce their rights, and often, those who choose to seek justice for online harms inflicted upon them are re-victimised in the processes they are made to endure by the justice system.[97]

Following Brenda Cossman's lead, if we adopt a reparative reading practice, and take seriously the opposing feminist perspectives briefly sketched out in the preceding two sections, what is the role of law in combating online gender-based violence?

Firstly, our own research findings and a growing body of evidence demonstrate that efforts to use the law to police online speech based only on its locutionary content, that is, based on the literal meaning of the words or their connotations, are futile and doomed to fail. This is because the whole gamut of behaviours which constitute gender-based violence cannot conceivably be enumerated or marked out in any meaningful way.

In order to appreciate the structural harms of pervasive online trolling and misogyny in ways that are relevant to law, we need to attend to the illocutionary force of such speech in systematically legitimating the discriminatory treatment of women. This, as argued above, is the contemporary context of virality emanating from the gaming of sociality by corporations for profit.

Second, if such laws that seek to criminalise sexual speech are enacted, they are likely to have unintended consequences and disproportionately criminalise non-normative sexual expression and reify gender and sexual hierarchies. Any legal intervention to tackle online gender-based violence must, therefore, be conversant with these concerns.

We suggest that the way to think about the regulation of online misogynistic speech on social media is a move away from criminal, carceral, and retributive notions of justice towards those based on a model of accountability, that foreground the effective delivery of justice in ways that are responsive to the needs of the victim.

To reiterate, our analysis in this section is limited to an examination of legal responses that would be effective in tackling the issue of online gender trolling and other speech-based forms of violence. There are other more extreme and egregious forms of technology-facilitated gender-based violence that certainly should invite criminal action, but this larger analysis remains out of the scope of this study.

In the next section, we consider the kinds of legal-institutional responses and platform enforcement mechanisms that can respond to online misogynistic speech in ways that keep pace with contemporary forms of speech regulation on social media.

5.2 Platform accountability

This study has demonstrated just how variable the forms of abuse faced by women on Twitter can be. We have seen instances of hate speech that are based on insinuations, dog whistles, and prejudicial stereotypes, not to mention those in which trolls deliberately employ tactics such as the use of special characters or even rhyme to evade automated detection. (See, for example, some of the tactics described in section 4.2.2). Any legal response, therefore, must be cognizant of the impossibility of effectively and exhaustively policing offensive speech based on its locutionary content.

In the landmark 2015 Shreya Singhal case, the Supreme Court held that a social media intermediary would be obligated to take down content only on receiving an order from a court or government authority, that is, only upon receiving "actual knowledge" asking it to take down the infringing material. The judgement is based on a delicately balanced set of considerations that seek to safeguard freedom of expression and prevent this right from being undermined in online environments.

What this research reveals is that the de facto enjoyment of the right to free expression online is not available to women, as it comes attached with disproportionate burdens. This points to the need to steer the regulatory debate beyond speech-related discussions about free speech towards an attention economy-focussed approach. Tech policy cannot be held hostage to irresolvable discussions regarding where to draw the line in the sand in regulating user-generated speech, while trolls are allowed free rein in the ongoing interregnum to indulge in abusive behaviour.

Moving past an individualistic frame of the victim-perpetrator binary toward a model of accountability which holds platforms responsible for the hostile and abusive online environments they foster and profit from is critical. A central component of this argument lies in dismantling the notion of the platform as a passive intermediary or a 'dumb conduit.' This involves emphasising the extent to which platforms already mediate content, and thereby regulate speech, not by directly muzzling or stifling user speech, but by steering and manipulating the attention of users in particular ways through algorithmic means—organising, ranking, recommending, hiding, and curating.

In an era characterised by information overload,[98] the crucial function of discerning what content is (and equally, what is not) deemed important to the public conversation, is not —as the platforms might have us believe— value (or content) neutral. It is informed by a capitalist, self-serving ethic that instrumentalises gender and social power relations to promote the kind of misogynistic behaviour we have reported on. As Zeynep Tufecki argues,

> *"The most effective forms of censorship today involve meddling with trust and attention, not muzzling speech itself. As a result, they don't look much like the old forms of censorship at all. They look like viral or coordinated harassment campaigns which harness the dynamics of viral outrage to impose an unbearable and disproportionate cost on the act of speaking out."*[99]

Once we disabuse ourselves of the false notion that the curation and recommendation function performed by platforms is somehow politically neutral, and recognise that these algorithms are developed and implemented with a narrow set of profit-oriented ends in mind, we can then propose other moral alternatives that are worth striving toward.

Further, as discussed at length in this report, the issue at the heart of online trolling and misogyny is not only the content of abusive speech, but also the volume and frequency of such messages that contribute to their potency and toxicity. The architecture of social media platforms, by quantifying the success of a message through metrics such as the engagement rate or meaningful social interaction (MSI),[100] prioritises shallow and instantaneous engagement by an ever-multiplying number of users, rather than sustained societal reflection on issues of immense social and political significance. When all engagement is treated equally by platforms (with respect to content), negative, sensationalist, divisive, simplistic, and abusive content is invariably amplified more than 'ordinary' social interactions and balanced or nuanced opinions.

This system is also easily exploited for gain by bad faith actors who, as we have discussed in section 4.1.1, indulge in deliberate acts to algorithmically ensure the spread of their messages.

Platforms must, therefore, be bound by law to implement ways of arresting the algorithmic amplification of misogynistic content as an ongoing commitment to their statutory duty of care owed to users. The following two subsections on the amplification of content and on transparency reporting highlight the key themes under which we make recommendations to ensure that platforms are made accountable for their users.

5.2.1 Amplification and virality of content

Recent revelations by Frances Haugen, a former Facebook data scientist turned whistleblower, provide evidence of misplaced priorities and the wilful disregard by Facebook toward matters that concern the health of public conversation. One set of leaked documents provides details of the Cross Check system, by which Facebook gave certain high-profile users special treatment by exempting them from the platform rules.[101] In a case involving the popular Brazilian footballer Neymar Jr., who was among those in the whitelisted category of users under Project Cross Check, Neymar responded to allegations of rape by publicly revealing the identity of the accuser, their private correspondence, and nude photos of the woman. According to a report in The Guardian, "[A]n internal review of the Neymar posts found that the video was viewed 56m times on Facebook and Instagram before its removal."[102] Under Facebook's rules, the normal procedure for users who post unauthorised nude photos is the deletion of the account. In Neymar's case, this rule was not followed even after the matter was escalated to company leadership. His account has since remained active.

Project Cross Check is an instructive example of how content moderation for prominent and influential personalities ought not to be done. Our findings also show that there are often cases where targeted trolling faced by women in public-political life is triggered by prominent users instigating pile-ons by hordes of trigger-happy trolls. Thus, rather than excluding users who have a large following from any kind of review, a more sensible strategy would be to impose greater responsibility on platforms to review and moderate content created or shared by accounts with greater visibility and a high number of followers (such as blue-tick accounts on Twitter).[103]

There are also preemptory measures that platforms can take to ensure that misogynistic content does not spread virally. One of the main routes by which content is made to 'go viral' on social media platforms is through the reshare/repost/retweet functionality. Research suggests that retweet or reshare cascades can be predicted with some degree of accuracy.[104] Platforms must invest in these capabilities and use them to arrest the viral spread of misogynistic content. In cases where predictions about an evolving cascade shape point to a potentially viral spread of content, platforms must take greater responsibility to ensure that it is not offensive or abusive.

Another set of preemptory measures revolve around the idea of "slowcial media," or slowing down the speed of interaction on social media and encouraging individual users to reflect on the content they post. Twitter has recently rolled out a new feature that will show "a self-moderation prompt to users who compose replies that the platform's algorithms recognise to be abusive."[105] Facebook's leaked internal research has also shown the centrality of the reshare function in spreading viral hate and disinformation.[106] Another proposal has been to get rid of the reshare functionality altogether to tackle the problems associated with speed and instantaneity on social media.

5.2.2 Transparency reporting

In recent times, the call for increased transparency from social media platforms has turned into something of a catch-all without much substantive meaning. Platforms themselves make use of this ambiguity by making magnanimous pronouncements of "committing to greater transparency" as just another means of public visibility management. Regular and comprehensive transparency reporting, however, is absolutely indispensable for any effective platform governance strategy. In order to get a clearer sense about the scale of online gender-based violence, it is essential that data regarding content takedowns, account suspensions, appeals against takedown or suspension decisions, number of grievances raised, and number of grievances resolved are released in public interest.[107]

The Santa Clara Principles on transparency and accountability in content moderation outline a bare minimum of best practices that platforms should adhere to in their transparency reporting.[108]

These three principles relate to reporting the number of posts/accounts actioned, issuing notices to affected users, and providing appeal mechanisms. At the time of writing, the reports submitted by SSMIs in India to comply with Rule 4(1)(d) of the IT Rules, 2021,[109] show that they are a long way from adhering to these principles.[110] Social media platforms in India are not responsive to user complaints, and the manner in which they disclose data about content decisions leaves much to be desired, as we outline below.

There is no consistency in the format of reporting across different platforms. In Facebook's compliance report for March 2022, the numbers on content actioned are presented under the community standards categories of hate speech, bullying and harassment, violence and incitement, nudity and sexual activity, and dangerous individuals/organisations. Google and WhatsApp's compliance reports, however, do not segregate this data at all and only provide aggregate figures. Such aggregated data is of little use in understanding the scale of a specific problem such as online gender-based violence. For this data to be actionable, it is crucial that platforms provide it in standardised formats and disaggregate it into relevant categories or policy areas.

Two pieces of evidence from the SSMI compliance reports for March 2022 may be of special import for our purposes. First, Twitter's compliance report states that the maximum number of complaints the company received related to abuse/ harassment. This confirms our own anecdotal data, collected as part of this research, on the pervasiveness of abuse and trolling on the platform. Second, Facebook's compliance report for the same period states that the "proactive rate" (the percentage of "content actioned" that the company detected proactively prior to any user reports) for content under the bullying and harassment category is significantly lower than categories such as nudity, graphic content, or child endangerment. Therefore, this means that maximum user complaints from India were received under the category of bullying and harassment, and that social media platforms are consistently failing to address the issue.[111]

As we have repeatedly asserted throughout this report, forms of trolling and harassment do not present themselves in easily recognisable ways through the use of explicit language or violent invectives. They take the form of seemingly harmless jokes, insinuations, and stereotypes that are often specific to socio-historical contexts, and derive potency from their volume and frequency. This is why proactive detection methods such as keyword filtering or image detection are not always effective in picking up on bullying and harassment. To counter this problem, it is crucial that platforms invest in training local human moderators who are equipped with the requisite cultural competence to be able to identify these forms of trolling, and augment proactive monitoring tools.

There is now a growing consensus that platforms need to go beyond the bare minimums of the Santa Clara Principles on transparency reporting, and provide disclosures about a broader range of issues.[112] One area where greater transparency is needed is in the use of algorithmic recommendation systems.[113] This entails giving users greater control over what they see on their own feeds in order to guard against the dangers of microtargeting, content optimisation, and troublingly, behavioural experiments conducted by platforms themselves.[114] Baking-in increased autonomy for users into the design of platforms is especially important in the regulatory model we outline in this report, in which platforms are the first port of call for addressing online gender-based violence.

Research conducted by Ben Wagner et al. comparing the transparency reporting of Twitter and Facebook in compliance with Germany's Network Enforcement Act (NetzDG) shows that there are significant variations in the way platforms use dark patterns (hidden interface design practices that subtly manipulate user behaviour through prompts and nudges) in the processes for raising a grievance. In this, the research highlights the importance of platform design practices, and the fact that transparency, "concerns not only providing data, but also how the visibility of the data […] is managed, by deciding how the data is provided and is framed."[115]

An important part of making social media platforms safer for women is to provide clear, easy, and quick ways to submit grievances. This kind of transparency can enable more informed and empowering decision-making for users. In addition, platforms should also be required to publicly disclose information to the public about advertisements (who bought them, who are the target audience, how much did they pay), bot activity (how many fake accounts are on the platform), and anonymised data about non-private user activity (which can be used by social science researchers).

5.3 Liability regimes for oversight of platform governance

It is clear that self-regulation has not yielded the desired results of making platforms safe spaces for women and marginalised groups, and that the hard edge of legal enforcement mechanisms and coercive action is needed. What is required is a legislation to hold platforms liable for online harms (with differential compliance obligations depending on the size of the platform) and the creation of mechanisms to oversee platform governance.

One route could be to set up an independent and autonomous regulatory authority with considerable discretionary powers to oversee social media regulation in India and ensure that platforms' terms of service are enforced uniformly and fairly. The UK, in its Online Harms White Paper of 2020, sets out a comparable regulatory framework to tackle a range of online harms through the appointment of an independent regulatory authority empowered with a range of enforcement powers to ensure that companies fulfil their statutory duty of care. These include powers "that would enable the regulator to disrupt the business activities of a non-compliant company, measures to impose liability on individual members of senior management, and measures to block non-compliant services."[116]

The moderation of dangerous, hateful, and misogynistic content by platforms would fall under the remit of this proposed regulatory authority. This would allow online gender-based violence on social media platforms to be tackled through a combination of compliance and deterrence-based regulatory strategies.

A central problem in content governance practices across platforms is that there

is little consistency in either identification or enforcement action when it comes to content flagged under specific policy areas such as hate speech or disinformation. As entities that hold enormous gatekeeping powers over speech, social media platforms must be held to rigorous legal standards. There is a need to establish a degree of normative coherence to ensure that platforms' terms of service are enforced predictably, fairly, and without arbitrariness.[117]
To this end, we lay out some overarching principles that can help to modulate decision-making by the regulatory authority.

- Content governance must be bound by a set of guiding principles on specific policy areas in order to advance the principle of international comity and harmonise adjudication practices across platforms.[118] While there will certainly be challenges in striking a balance between the poles of universalism and relativism in such an approach, there is at the very least, the need for global agreement on some basic principles of content governance on online platforms. Platforms must also use terminologies in a consistent manner and make efforts to reach a convergence on their terms of service/community standards. To ensure that artificial intelligence (AI) identification/proactive monitoring meets the standard of normative coherence, tools used must not only analyse content substantively (by looking at keywords or images), but also according to their level of dissemination and virality.[119]

- Especially relevant in the context of tackling online sexist hate, a gendered perspective on the right to freedom of expression must include a framework for promoting positive freedoms (freedom for), such as the right to public participation, or the right to publicness, as well as negative freedoms (freedom from). Under the prevailing legal position, negative freedoms largely animate moderation policies and platform design.[120]

- Any measures to restrict gender-based hate 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. As a point of reference, Germany's NetzDG prescribes different response times for manifestly unlawful content (24 hours) and other illegal content (seven days after receipt of complaint). The regulatory authority should not be empowered to initiate criminal proceedings, except in the most egregious cases of harm.

- Comprehensive and regular transparency reporting is essential to accurately diagnose and address the problem of online gender-based violence. In addition to takedown decisions, disclosures must improve transparency in algorithm design and implementation. Transparency reporting formats must be standardised across platforms and provide disaggregated data on gender-based violence according to specific policy areas.

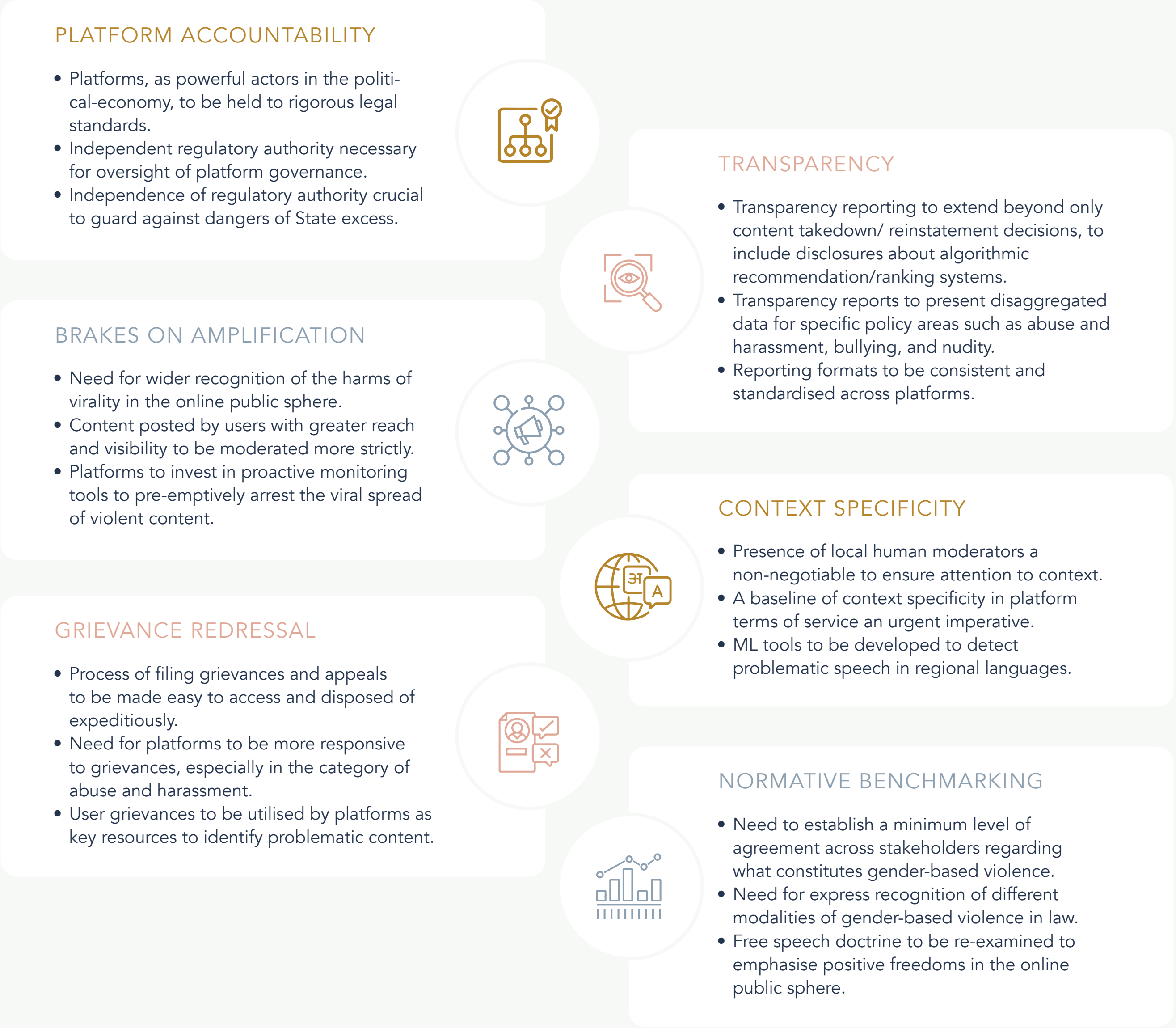
- Platforms must put in place responsive grievance redressal mechanisms for users to submit their complaints. Robust appeal mechanisms are also crucial to offset the significant risk of wrongful decisions by platforms. Both grievance redressal and appellate mechanisms should be easily and quickly accessible 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 should be tasked with investigating practices such as shadowbanning and use of dark patterns by platforms, as well as ensuring that the trending page on a platform does not cause hateful content to spread virally.

Based on all of the above considerations, Infographic 10 provides a summary of the recommendations we make for platform governance. This study has sought to shed light on the pervasive problem of online gender-based trolling.

By analysing a small sample of abusive tweets directed at women in public-political life on Twitter in India, the study traces its entanglements with the ideological apparatus of Brahminical patriarchy, unpacking the immense gravity of the problem. Arguing against the contention that gender trolling is a mild form of abuse, it demonstrates, instead, that online misogyny contributes to legitimising a regressive view of women's place in society.

Through the recommendations, the report makes the case for legal-institutional responses that are based on the tenet of platform accountability, rather than the individualistic victim-perpetrator binary of criminal law.



PLATFORM ACCOUNTABILITY

- Platforms, as powerful actors in the political-economy, to be held to rigorous legal standards.
- Independent regulatory authority necessary for oversight of platform governance.
- Independence of regulatory authority crucial to guard against dangers of State excess.

BRAKES ON AMPLIFICATION

- Need for wider recognition of the harms of virality in the online public sphere.
- Content posted by users with greater reach and visibility to be moderated more strictly.
- Platforms to invest in proactive monitoring tools to pre-emptively arrest the viral spread of violent content.

GRIEVANCE REDRESSAL

- Process of filing grievances and appeals to be made easy to access and disposed of expeditiously.
- Need for platforms to be more responsive to grievances, especially in the category of abuse and harassment.
- User grievances to be utilised by platforms as key resources to identify problematic content.

TRANSPARENCY

- Transparency reporting to extend beyond only content takedown/ reinstatement decisions, to include disclosures about algorithmic recommendation/ranking systems.
- Transparency reports to present disaggregated data for specific policy areas such as abuse and harassment, bullying, and nudity.
- Reporting formats to be consistent and standardised across platforms.

CONTEXT SPECIFICITY

- Presence of local human moderators a non-negotiable to ensure attention to context.
- A baseline of context specificity in platform terms of service an urgent imperative.
- ML tools to be developed to detect problematic speech in regional languages.

NORMATIVE BENCHMARKING

- Need to establish a minimum level of agreement across stakeholders regarding what constitutes gender-based violence.
- Need for express recognition of different modalities of gender-based violence in law.
- Free speech doctrine to be re-examined to emphasise positive freedoms in the online public sphere.

Infographic 10: Recommendations targeted at platforms in the above areas

ENDNOTES

[1] Cohen, J. (2019). Between Truth and Power: The Legal Constructions of Information Capitalism. Oxford University Press; see also, Sinha, A., & Basu, A. (2021, August 13). Why Metaphors for Data Matter. Bot Populi. botpopuli.net

[2] Narrain, S. (2020). From the Rhetorical Software to the 'Hardware of the Law': Regulating Hate Speech Online in India. GNLU Law & Society Review. academia.edu

[3] Das, V. (2008, June). Violence, Gender and Subjectivity. Annual Review of Anthropology, 37(1), 284.

[4] Pal, J., & Gonawela, A. (2017). Studying Political Communication on Twitter: The Case for Small Data. Current Opinion in Behavioral Sciences, 18:97–102. joyojeet.people.si.umich.edu

[5] Tukde Tukde Gang is a pejorative political catchphrase used in India by the right-wing establishment and its sympathisers, accusing their critics of allegedly supporting sedition and secessionism. See: en.wikipedia.org

[6] Jamil, G. (2021, September 3). India's 'love jihad' anti-conversion laws aim to further oppress minorities, and it's working. The Conversation. theconversation.com

[7] Chakravarti, U. (1993, April 3). Conceptualising Brahmanical Patriarchy in Early India: Gender, Caste, Class and State. Economic and Political Weekly, 28 (14).

[8] id

[9] Das, V. (1995). National Honour and Practical Kinship: Of Unwanted Women and Children. In Critical Events. Oxford University Press.

[10] Abducted Persons (Recovery and Restoration) Act, 1949. Available at: indiacode.nic.in

[11] The Quint. (2021, July 1). Srinagar 'Conversion' Row: 'What About Our Agency,' Tweet Women. thequint.com

[12] Arya, S. (October, 2020). Theorising Gender in South Asia: Dalit Feminist Perspective. CASTE: A Global Journal on Social Exclusion, 1 (2,), XI–XXIV. journals.library.brandeis.edu

[13] Munusamy, K. (2018, June 7). Intersection of identities: Online gender and caste-based violence. GenderIT.org. genderit.org

[14] Lochan, V. (2020, July 7). Casteism In Our Words: 10 Casteist Slurs And Why We Need To Stop Throwing Them Around. Homegrown. homegrown.co.in

[15] Dube, M. (2015, September 6). Jootha is just untouchability by another name. Scroll. scroll.in

[16] For a more representative list of connotations, see **नीच**, Shabdkosh English Hindi Dictionary, shabdkosh.com

[17] "Encounter killing" is a term used in India, Pakistan, Bangladesh, and Sri Lanka since the late 20th century to describe extrajudicial killings by the police or the armed forces, supposedly in self-defence, when they encounter suspected gangsters or terrorists in a shootout situation. See: en.wikipedia.org

[18] Sajlan, D. (2021). Hate Speech against Dalits on Social Media: Would a Penny Sparrow be Prosecuted in India for Online Hate Speech? CASTE / A Global Journal on Social Exclusion, 2(1), 77-96. journals.library.brandeis.edu

[19] Deshpande, S. (2013). Caste and Castelessness in the Indian Republic: Towards a Biography of the 'General Category'. Review of Development & Change, 18(1), 3-18.

[20] McConnell-Ginet, S. (2012). Linguistics and Gender Studies. Handbook of the Philosophy of Science, 503-530. sciencedirect.com

[21] For similar findings about the indirectness of online casteist speech, see also: Kain, D., Narayan, S., Sarkar, T., & Grover, G. (2021). Online caste-hate speech: Pervasive discrimination and humiliation on social media. The Center for Internet and Society. cis-india.org

[22] Shanmugavelan, M. (2021). Caste-hate speech: Addressing hate speech based on work and descent. International Dalit Solidarity Network (IDSN). idsn.org

[23] Apoorvanand. (2021, October 12). The Banality of India's Islamophobia. The India Forum. theindiaforum.in

[24] The Sangh Parivar refers, as an umbrella term, to the collection of Hindu nationalist organisations spawned by the Rashtriya Swayamsevak Sangh (RSS). See: en.wikipedia.org

[25] Soundararajan, T., Kumar, A., Nair, P., & Greely, J. (2019). Facebook India: Towards The Tipping Point of Violence Caste and Religious Hate Speech. Equality Labs, USA. See: equalitylabs.org

[26] See note 5

[27] Nyst, C., & Monaco, N. (2018). State-Sponsored Trolling: How Governments Are Deploying Disinformation as Part of Broader Digital Harassment Campaigns. Institute for the Future. iftf.org

[28] Saha, P., Mathew, B., Garimella. K., & Mukherjee, A. (2021, April). Short is the Road that Leads from Fear to Hate: Fear Speech in Indian WhatsApp Groups. Proceedings of the Web Conference. dl.acm.org

[29] Afsaruddin, A. (2013, September). Striving in the Path of God: Jihad and Martyrdom in Islamic Thought. Oxford Scholarship Online. oxford.universitypressscholarship.com

[30] supra note 6.

[31] Jafri, A., & Aafaq, Z. (2021, May 21). Unchecked Tsunami Of Online Sexual Violence By Hindu Right Against India's Muslim Women. Article 14. article-14.com

[32] Salim, M. (2022, January 16). 'Bulli Bai', 'Sulli Deals': On Being Put Up for 'Auction' as an Indian Muslim Woman. The Wire. thewire.in

[33] IWMBuzz. (2019, April 14). The most iconic vamp of television "Komolika" From Kasautii Zindagii Kay. IWMBuzz. iwmbuzz.com

[34] Mohan, S. (2021, October 23). My Bindi Does Not Speak My Culture – The Burden of Propriety on Indian Women. The Quint. thequint.com

[35] Mushrif, M. (2021, August 11). The Comeback Of Blatant Misogyny This Olympic Season. Feminism in India. feminisminindia.com

[36] Udupa, S. (2018). Gaali cultures: The politics of abusive exchange on social media. New

Media & Society, 20(4), 1506-1522.

37 Kaur, R., & Mazzarella, W. (2009). Censorship in South Asia: Cultural Regulation from Sedition to Seduction. In Between Sedition and Seduction: Thinking Censorship in South Asia.

38 Geybulla, A. (2016, November 22). In the crosshairs of Azerbaijan's patriotic trolls. Open Democracy. opendemocracy.net

39 Merriam-Webster. Troll. In Merriam-Webster's online dictionary. merriam-webster.com

40 Chakravartty, P., & Roy, S. (2015). Mr. Modi Goes to Delhi: Mediated Populism and the 2014 Indian Elections. Television and New Media, 16 (4). journals.sagepub.com

41 Bandurski, D. (2008, September 24). China's Guerrilla War for the Web. Home is Where the Heart Dwells. blogs.harvard.edu

42 supra note 36 at pp.1509.

43 id.

44 Lulz, a corruption of LOL, online shorthand for laugh-out-loud, is defined as "fun, laughter, or amusement, especially that derived at another's expense". See: Lexico English Dictionary. Lulz. lexico.com

45 Phillips, W., & Milner, R.M. (2021). You Are Here: A Field Guide for Navigating Polarized Speech, Conspiracy Theories, and Our Polluted Media Landscape. (pp. 78). MIT Press.

46 Kapur, R. (1996, April 20).Who Draws the Line-Feminist Reflections on Speech and Censorship. Economic and Political Weekly, 31, 16-17. epw.in

47 supra note 37, at pp. 6.

48 Swara Bhasker. [@ReallySwara]. (2021, August 20). Status. [Twitter Profile]. Twitter. Retrieved from twitter.com

49 Nguyen, C.T. (2021). How Twitter gamifies communication. In Jennifer Lackey (ed.), Applied Epistemology. (pp. 410-436). Oxford University Press.

50 RoyIntPhilosophy. (2019, December 5). The Gamification of Public Discourse. Royal Institute of Philosophy. [Video]. YouTube. youtube.com

51 ITforChange. (2021, February). How Can We Hold Social Media Accountable for Misogyny? I Sexism and the Online Publics I Session 2. [Video]. YouTube.youtube.com at 13:00.

52 Curlew, A. E. (2019). Undisciplined Performativity: A Sociological Approach to Anonymity. Social Media + Society. journals.sagepub.com

53 Walsh, M.J., & Baker, S.A. (2021, August 31). Twitter's design stokes hostility and controversy. Here's why, and how it might change. The Conversation. theconversation.com

54 A research study conducted by IT for Change in three South Indian states showed that young women used similar aviodance strategies to navigate online spaces. See Gurumurthy, A., Vasudevan, A., & Chami, N. (2019). Born digital, Born free? A socio-legal study on young women's experiences of online violence in South India. IT for Change.itforchange.net

55 supra note 53.

56 Spangles, T. (2016, July). Twitter CEO Jack Dorsey: 'We Need to Do Better' at Curbing User-Targeted Abuse. Variety. variety.com

[57] Word we're Watching: Ratioed. Merriam Webster Dictionary. merriam-webster.com

[58] Mateus, S. (2022, January). Publicness beyond the public sphere. Mediapolis Revista de Comunicação Jornalismo e Espaço Público. researchgate.net

[59] Splichal, S. (2018). Publicness–Privateness: The Liquefaction of "The Great Dichotomy". Javnost - The Public. tandfonline.com

[60] See, for instance, the concept of illocutionary disablement in Silencing Acts: The Law and Illocutionary Disablement. (2019, May 12). Law and the Humanities LLM. blogs.kent.ac.uk

[61] Lim, M. Algorithmic enclaves: Affective politics and algorithms in the neoliberal social media landscape. In Boler, M., & Davis, E. (eds.), Affective Politics of Digital Media: Propaganda by Other Means (pp. 186-203). New York & London: Routledge, 2020

[62] Hansen, T.B.(2021). The Law of Force: The Violent Heart of Indian Politics. (pp. 28). Aleph Book Company.

[63] See, for instance: Chowdhury, R. (2021, October 21). Examining algorithmic amplification of political content on Twitter. Twitter Blog. blog.twitter.com

[64] Udupa, S. (2014, February 15). Aam Aadmi: Decoding the Media Logics. Economic and Political Weekly, 49 (7). epw.in; See also, Banerjee, M. (2022, February 3). Rajneeti Or Politics? Outlook. outlookindia.com

[65] Bhatia, K.V. (2021). Religious Subjectivities and Digital Collectivities on Social Networking Sites in India. Studies in Indian Politics, 9 (1). journals.sagepub.com

[66] The Overton window is the range of policies politically acceptable to the mainstream population at a given time. It is also known as the window of discourse. See: en.wikipedia.org

[67] 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). (pp 73).

[68] Chaturvedi, S. (2016). I am a Troll: Inside the Secret World of the BJP's Digital Army. Juggernaut.

[69] supra note 67.

[70] Kohli, K. (2017, August 31). Templated Tweets, Trolls Deployed to Proclaim 'Success' of Demonetisation on Social Media. The Wire. thewire.in

[71] Kappan, R. (2020, January 5). Free Netflix, sex chats: Callers lured to 'support CAA'. Deccan Herald. deccanherald.com

[72] Lee, R. (2019). Extreme Speechl Extreme Speech in Myanmar: The Role of State Media in the Rohingya Forced Migration Crisis. International Journal of Communication. ijoc.org

[73] Hakopian, M. (2021, October 21). Algolinguicism: Translating Language Justice to Digital Platforms. A New AI Lexicon. Retrieved from medium.com

[74] Culliford, E., & Heath, B. (2021, October 21). Language Gaps in Facebook's Content Moderation System Allowed Abusive Posts on Platform: Report. The Wire. thewire.in

[75] Gillespie, T. (2018). Custodians of the Internet: platforms, content moderation, and the hidden decisions that shape social media. (pp. 59). Yale University Press.

[76] Caplan, R. (2018, November). Content or Context Moderation? Artisanal, Community-reliant

and Industrial Approaches. Data & Society. datasociety.net

77 SSMI compliance reports for this period are accessible here: internetfreedom.in

78 NWMI demands action against online abuse of journalist Kavin Malar. (2020, September 1). Network of Women in Media, India. nwmindia.org

79 Raghavan, A. (2021). The Internet-Enabled Assault on Women's Democratic Rights and Freedoms. IT for Change. itforchange.net

80 Cossman, B. (2021). The New Sex Wars: Sexual Harm in the #MeToo Era. NYU Press.

81 Cossman, B. (2018, September). #MeToo, Sex Wars 2.0 and the Power of Law. Asian Yearbook of Human Rights and Humanitarian Law. ssrn.com

82 Menon, N. (2019, February 28). How the Feminist Conversation Around Sexual Harassment Has Evolved. The Wire. thewire.in; See also, Meghana. (2018, March 13). A Practitioner Of Finger-Tip Activism Responds To Nivedita Menon. Feminism in India. feminisminindia.com

83 supra note 81 at pp 16.

84 Sedgwick, E. (2022). Paranoid and Reparative Readin, or, You're so Paranoid, You Probably Think This Essay is About You. In Touching, Feeling: Affect, Pedagogy. Performativity. Duke University Press

85 Austin, J.L. (1962). How to Do Things With Words. Oxford, Clarendon Press.

86 Langton, R. (1993). Speech Acts and Unspeakable Acts. Philosophy & Public Affairs, 22 (4), 293-330.

87 Consider the debates about the proliferation of rape myths, and Robin Morgan's famous phrase, "Pornography is the theory, and rape is the practice." From, Going Too Far: The Personal Chronicle of a Feminist. (1978). Vintage Books USA.

88 supra note 79 at pp. 20.

89 De Silva, A. (2020). Addressing the Vilification of Women: A Functional Theory of Harm and Implications for Law. Melbourne University Law Review, 43 (3), 995.

90 See for example, the "Denim Day" campaign to raise awareness about sexual assault, which came about from the Italian SC judgement which debunked common rape myths. denimdayinfo.org

91 Legal Challenge to the Kerala Police Act Amendment. (2020, November 24). Software Freedom Law Center, India. sflc.in

92 Explained: Why Kerala Police Act Amendment is controversial. (2020, November 23). Money Control. moneycontrol.com

93 For more examples in the Indian context from the recent past, see also, Bailey, R., & and Bhandari, V. (2021). Towards Holistic Regulation of Online Hate Speech. IT for Change. itforchange.net. The authors cite cases involving the journalists Mohammed Zubair and Prashant Kanojia as evidence of the misuse of sections 67 and 67A of the IT Act which are ostensibly intended to protect women from online sexual harms.

94 Hooton, C. (2014, December 2). A long list of sex acts just got banned in UK porn. Independent. independent.co.uk

95 Srinivasan, A. (2021). The Right to Sex: Feminism in the Twenty-First Century. (pp. 58). Farrar,

Straus and Giroux.

96 Datta, B., et al. (2018). Guavas and Genitals: A research study in Section 67 of the Information Technology Act. IT for Change. itforchange.net

97 Salim, M. (2021). How Women from Marginalised Communities Navigate Online Gendered Hate and Violence. IT for Change. itforchange.net

98 Andrejevic, M. (2013). Infoglut: How Too Much Information Is Changing the Way We Think and Know. Routledge.

99 Tufecki, Z. (2020, October 28). The Problem With (All) The Tech Hearings. The Insight. theinsight.org

100 Hagey, K., & Horowitz, J. (2021, September 15). Facebook Tried to Make Its Platform a Healthier Place. It Got Angrier Instead. The Wall Street Journal. wsj.com

101 Horowitz, J. (2021, September 13). Facebook Says Its Rules Apply to All. Company Documents Reveal a Secret Elite That's Exempt. The Wall Street Journal. wsj.com

102 Milmo, D. (2021, September 13). Facebook: some high-profile users 'allowed to break platform's rules'. The Guardian. theguardian.com

103 supra note 79 at pp. 28.

104 Cheng, J., et al. (2014, April). Can cascades be predicted?. WWW '14: Proceedings of the 23rd international conference on World Wide Web. (pp. 925–936). dl.acm.org

105 Graff, M. (2021, May 22). Twitter is taking a step towards 'slowcial media' by asking users to rethink abusive messages. Scroll. scroll.in

106 Hagey, K., & Horwitz, J. The Journal. [Audio]. Gimlet Media and The Wall Street Journal. wsj.com

107 MacCarthy, M. (2022, May 5). Transparency Is the Best First Step towards Better Digital Governance. Centre for International Governance Innovation. cigionline.org

108 The Santa Clara Principles On Transparency and Accountability in Content Moderation. santaclaraprinciples.org

109 Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021. meity.gov

110 #SocialMediaComplianceWatch: Analysis of Social Media Compliance Reports for the month of March 2022. (2022, May 27). Internet Freedom Foundation. internetfreedom.in

111 #SocialMediaComplianceWatch: analysis of Social Media Compliance Reports of August, 2021. (2021, October 25). Internet Freedom Foundation. internetfreedom.in

112 Kornbluh, K., & Goodman, E.P. (2019). Bringing Truth to the Internet. Democracy Journal, 53. democracyjournal.org

113 Singh, S. (2020, March 25). Why Am I Seeing This? How Video and E-Commerce Platforms Use Recommendation Systems to Shape User Experiences. New America. newamerica.org

114 Meyer, R. (2014, June 29). Everything We Know About Facebook's Secret Mood-Manipulation Experiment. The Atlantic. theatlantic.com

115 Wagner, B., et al. (2020). Regulating Transparency? Facebook, Twitter and the German

Network Enforcement Act. ACM Conference on Fairness, Accountability, and Transparency. researchgate.net

116 Online Harms White Paper 2020. Government of UK. gov.uk

117 Chami, N., & Kanchan, T.(2021, March 24). A Feminist Social Media Future: How Do We Get There?. Bot Populi. botpopuli.net

118 See, IT for Change's submission to the Special Rapporteur on the right to freedom of opinion and expression: itforchange.net

119 Structuring Questions: Content moderation in relation to Covid-19. (2020, April 20). Internet and Jurisdiction Policy Network. internetjurisdiction.net

120 See the discussion on "communitarian ethics" in Phillips, W., & Milner, R.M. (2021). You Are Here: A Field Guide for Navigating Polarized Speech, Conspiracy Theories, and Our Polluted Media Landscape. MIT Press.

IT
for
Change

IDRC · CRDI
Canada