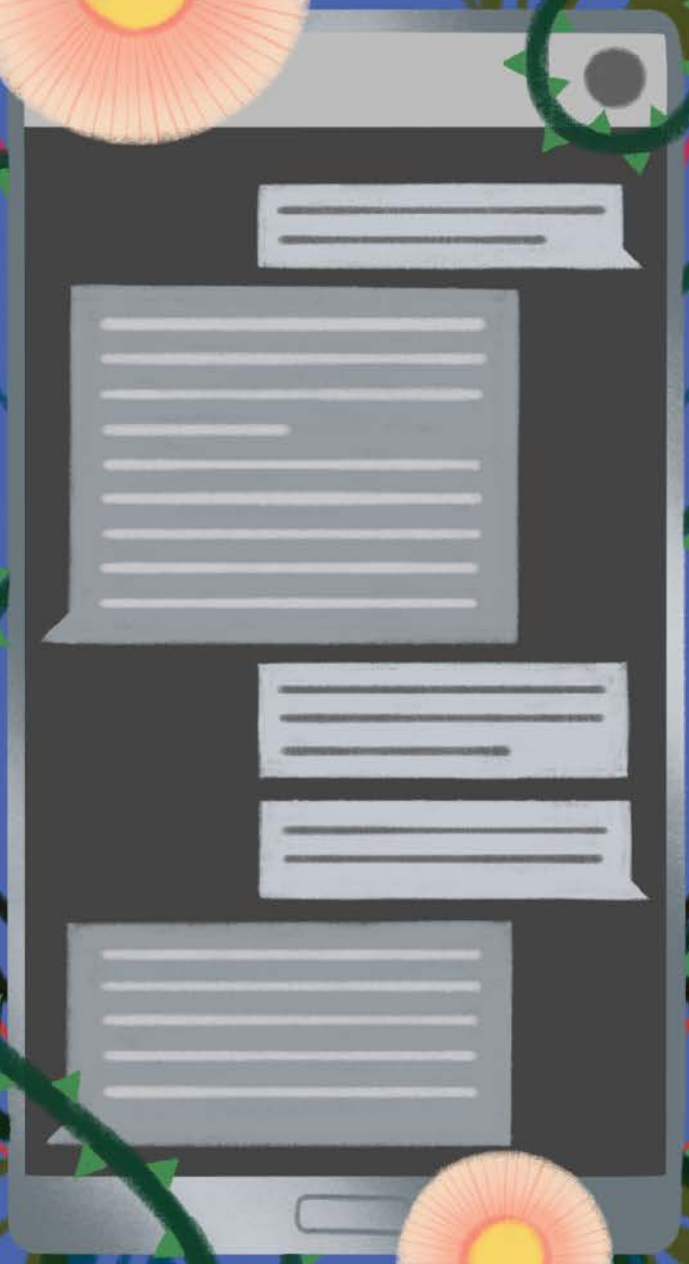
Online Caste-Hate Speech:

Pervasive Discrimination and Humiliation on Social Media



Online caste-hate speech: Pervasive discrimination and humiliation on social media

Published on 15 December 2021

Authors

Damni Kain, Shivangi Narayan, Torsha Sarkar and Gurshabad Grover

All authors contributed equally to the report

About CIS

The Centre for Internet and Society (CIS) is a non-profit organisation that undertakes interdisciplinary research on internet and digital technologies from policy and academic perspectives. Through its diverse initiatives, CIS explores, intervenes in, and advances contemporary discourse and regulatory practices around internet, technology, and society in India, and elsewhere.

Acknowledgements

This research was developed by the Centre for Internet and Society with support from the project ***Challenge: challenging hate narratives and violations of freedom of religion and expression online in Asia***, implemented by the Association for Progressive Communications (APC) with funds from the European Instrument for Democracy and Human Rights (EIDHR).

The authors are grateful to Beena Pallical, Srishty Ranjan, Christina Dhanraj, Manjula Pradeep, Sumeet Samos, Laxman Yadav, Jitendra Meena, Dipankar Kamble, Pardeep Singh Attri, Harishchandra Sukhdeve and Shafiullah Anis for taking out time to share their experience with and insights on online caste-hate speech.

The authors are deeply grateful to Dr. Murali Shanmugavelan for his ideas and detailed review of the report, and for writing its preface. The use of 'caste-hate speech' in the title and report owes credit to his work with the International Dalit Solidary Network.

The authors are thankful to Ambika Tandon and Cheshta Arora for their review and feedback; to Cheshta Arora for also assisting with the research and interviews.

The interviews described in this report were transcribed, thanks to the resourceful individuals at Team Sociolegalliterary. The design of the layout of the report was done by Abhilasha Prajapati, and the illustrations were done by Farah Ahmad. The report also benefited immensely from copyediting by the Clean Copy.

Any opinion or mistake in the report remains the authors'.

Disclosure: CIS is a recipient of research grants from Facebook India Pvt. Ltd.

This work is shared under the Creative Commons Attribution 4.0 International license.

Table of contents

Executive summary

In India, religious texts, social customs, rituals, and everyday cultural practices legitimise the use of hate speech against marginalised caste groups. Notions of 'purity' of "upper-caste" groups, and conversely of 'pollution' of "lower-caste" groups, have made the latter subject to discrimination, violence, and dehumanisation. These dynamics invariably manifest online, with social media platforms becoming sites of caste discrimination and humiliation.

This report explores two research questions. First, what are the specific contours of caste-hate speech and abuse online? Semi-structured interviews with 12 scholars and activists belonging to DBA groups show that marginalised groups regularly face hate and harassment based on their caste. In addition to the overt hate, DBA individuals and groups are often targeted with abuse for availing reservations – a constitutionally mandated right. More covert forms of hate and abuse are also prevalent: trolls mix caste names and words from different languages together so that their comments appear meaningless to individuals who are not keenly aware of the local context.

Such hateful expression often emerges as a reaction from "upper-caste" groups to DBA resistance and social justice movements.

Our respondents reported that the hateful expression can sometimes silence caste-marginalised groups and individuals, exclude them from conversations, and adversely impact their physical and mental wellbeing.

The second question we explore is how popular social media platforms and online spaces moderate caste-hate speech and abuse. We analysed the community guidelines, policies, and transparency reports of Facebook, Twitter, YouTube, and Clubhouse. We find that Facebook, Twitter, and Youtube incorporated 'caste' as a protected characteristic in their hate speech and harassment policies only in the last two or three years – many years after they entered Indian and South Asian markets — showing a disregard for the regional contexts of their users. Even after these policy changes, many platforms – whose forms for reporting harmful content list gender and race – still do not list caste.

Social media companies should radically increase their investment and capacity in understanding regional contexts and languages; they must focus on the dynamics of casteist hate and abuse. They will need to collaborate with a diverse set of DBA activists to ensure that their community guidelines effectively tackle overt, covert, and hyper-local forms of caste-hate speech and abuse, and that their implementation and reporting processes match these policy commitments.

Preface

*By **Murali Shanmugavelan***
Faculty Fellow – Race and Technology, Data and Society

Caste, perhaps one of the oldest surviving social hierarchies in the world, is the pervasive glue – or rupture, depending on one's caste – that overshadows all aspects of development in caste-affected countries. Caste is enacted through various cultural codes, and the penalty for breaching these codes, in Ambedkar's words, is ex-communication.

Caste affords social privileges to those who occupy the upper rungs of hierarchy and power. At the same time, for others – especially Dalits – it implies adverse experiences in everyday life, such as humiliation, separation, (forced) subordination, and degradation. These everyday experiences are rooted in rituals, social interactions, everyday conversations, signs, memes, infographics, and popular culture. In addition, speech and communication perpetuate and normalise caste-based hierarchies, while casteist speech humiliates and dehumanises Dalits, Bahujans, and Adivasis (DBAs). Therefore, wherever there is caste discrimination, there is caste-hate speech.

This report by the Centre for Internet and Society is a timely analysis of the severe problems posed by caste-hate speech – an issue often neglected by governments and private companies. While acknowledging the power of the internet, which has created a platform for many Dalits and members of oppressed groups, the report presents everyday experiences of caste-related online harms through a set of interviews with individuals with 'strong, assertive profiles' on social media platforms.

As one participant put it succinctly: Dalits attract hate speech for merely existing online. Mocking and humiliating people based on their caste could be "very damaging to [the] children and teens of social media" in the long run – it is a psychological issue that deserves policy attention so that we can build a safe and caste-sensitive internet.

In response to critiques that have pointed out the exclusion of concerns relating to race, gender and sexuality, technologists and communication policymakers have become more sensitive to the workings of power and its manifestations of domination. However, through a careful analysis of social media platforms, this report clearly makes the case that tech corporations continue to ignore caste as a protected category.

Caste-hate speech dehumanises, incites discrimination, and in its extreme form, incites physical violence – all clear violations of fundamental human rights. As caste discrimination is being recognised as a global phenomenon, tech corporations should become more sensitive towards the realities of caste, and protect Dalits and other oppressed caste groups from bullying and harassment.

Introduction

Whenever I present an argument regarding the rights of the marginalised on social media, I am not countered by a logical counter-argument (using texts [or] references), but I am humiliated because of the surname that I carry. This I see as caste-based hate speech, which makes even opining and discussion impossible. The quality and validity of my arguments are disregarded because hate speech overpowers the content of what I speak.
– *Dr. Laxman Yadav,Assistant Professor, University of Delhi*[1]

In an unequal society – such as that of South Asia – hate speech has developed out of unequal power relations, which determine one's 'vulnerability' to extreme forms of discrimination.[2] Hate speech is inflicted based on religion, gender, sexuality, disability, nationality, race, and caste.[3] Caste is a system of coerced hierarchy that subjects people to ascriptive identities based on the jati that they are born into. It is an all-pervasive phenomenon that manifests 'graded inequalities'[4] in the social, political, economic, cultural, and religious realms. Caste, which is accompanied by notions of 'pollution' of "lower caste" groups, is the underlying driver for humiliation, discrimination, and violence against them.[5] Caste discrimination affects hundreds of millions globally. In India alone, caste-oppressed groups include at least 300 million people.[6]

Such hate speech violates the dignity[7] of marginalised groups, leading to a significant negative impact on their social, economic, political, and mental wellbeing.[8] A common argument by free speech proponents against the regulation of hate speech is one of a 'marketplace of ideas' or of self-governance, suggesting that hate speech would naturally attract counter-speech and be self-correcting in a functioning democracy. However, the tangible presence of hate speech can have the effect of silencing exactly those at the forefront of expressing dissent against that hate speech.[9]

1. Interview with Dr. Laxman Yadav, July 2021
2. Vaghela, P., Mothilal, R. K., & Pal, J. (2020). Indian Political Twitter and Caste Discrimination—How Representation Does Not Equal Inclusion in Lok Sabha Networks. *ArXiv: 2007.15863* [Cs]. http://arxiv.org/abs/2007.15863.
3. Soundararajan, T., Kumar, A., Nair, P., Greely, J. (2019). *Facebook India - Towards a Tipping Point of Violence Caste and Religious Hate Speech.* Equality Labs. https://www.equalitylabs.org/facebookindiareport
4. Ambedkar, B. R., & Moon, V. (2014). *Dr. Babasaheb Ambedkar: Writings and speeches* (Vol. 3). Dr. Ambedkar Foundation. https://www.mea.gov.in/Images/attach/amb/Volume_03.pdf
5. Berg, D. E. (2020). *Dynamics of Caste and Law: Dalits, Oppression and Constitutional Democracy in India: 11 (First edition). Cambridge University Press.*
6. Soundararajan, 2019. Op.cit.
7. 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 https://doi.org/10.26812/caste.v2i1.260
8. Matsuda, M. J., Lawrence, C. R., Delgado, R., & Crenshaw, K. W. (1993). Words That Wound: Critical Race Theory, Assaultive Speech, And The First Amendment. Routledge & CRC Press.
9. Sellars, A. (2016). *Defining Hate Speech* (SSRN Scholarly Paper ID 2882244). Social Science Research Network. https://doi.org/10.2139/ssrn.2882244

These dynamics are invariably represented online. On the one hand, digital platforms are used to run awareness campaigns, social movements and online protests, and advance the cause of social justice. In particular, many anti-caste activists, academics, and leaders argue that digital platforms have provided a space for the marginalised to voice their opinions.[10] These voices ask for accountability, critique Brahmanic state policies, and challenge casteist acts.[11]

On the other hand, digital platforms remain sites of discrimination and humiliation, replicating – and at times exacerbating – offline caste discrimination. A 2019 report by Equality Labs noted how "Indian casteist hate speech is part of an ecosystem of violence designed to shame, intimidate, and keep caste oppressed communities from asserting their rights and participating as equals in society."[12] The report found many instances of caste-hate speech, including casteist slurs, messages against anti-caste leaders, and hatred of inter-caste relationships.

The response of social media companies to online caste-hate speech and abuse (OCHS) is often inadequate. Critics claim that Big Tech companies inconsistently invoke their community standards and ignore violations of human rights.

In many instances, social media platforms have implemented their policies in ways that shut down critics of the Hindu nationalism while simultaneously allowing the abuse of Dalits and Muslims.[13]

OCHS, even in its most overt forms, often goes unaddressed by digital platforms. There are also covert forms of discrimination that proliferate online. Some of the online casteist abuse in India can be best captured as gaali – a culture of abuse with blurred boundaries between comedy, insult, shame, and abuse.[14] Caste-hate speech can prima facie appear to some as mere 'comedy' or 'opinion' but be intended to insult deeply.

This report contributes to a broader understanding of caste-hate speech on digital platforms. We briefly discuss the existing literature on online caste-hate speech. We then draw insights from the analyses, opinions, and experiences of our respondents. We move on to analyse how four online platforms (Facebook, Twitter, Youtube, and Clubhouse) are equipped to deal with OCHS. Finally, we provide specific pointers and recommendations to these platforms, beyond basic content moderation, to counter caste-hate speech.

10. See Mitra, A. (2001). Marginal Voices in Cyberspace. New Media & Society, 3(1), 29–48. https://doi.org/10.1177/1461444801003001003
11. Ibid.
12. Soundararajan, 2019. Op.cit.
13. Chopra, P. C. & R. (2021, June 16). Can the subaltern tweet?: Holding Big Tech accountable on caste. *Scroll.In*. https://scroll.in/article/997105/can-the-subaltern-tweet-holding-big-tech-accountable-on-caste.
14. Udupa, S. (2018). Gaali Cultures: The Politics Of Abusive Exchange On Social Media. *New Media & Society*, 20(4), 1506–1522. https://doi.org/10.1177/1461444817698776.

Scope and Methodology

We explore two research questions in this report. First, how do caste-hate speech and abuse play out online? Second, how do popular social media platforms and online spaces moderate caste-hate speech and abuse?

First, we summarise significant scholarship on caste-hate speech. We then use a qualitative methodology to arrive at a descriptive understanding of online caste-hate speech and abuse, including the specificities of how it was experienced and resisted. We conducted 12 interviews with respondents belonging to DBA groups. These respondents – chosen using purposive sampling – are intellectuals, activists, and individuals who are on one or more online social media platforms. We used semi-structured interviews to afford respondents the space to explain their experiences of dealing with OCHS. An indicative list of the questions that we posed to the respondents is available in Annexure 1

We spoke to 12 respondents, 10 of whom were Dalits. These 10 respondents included one Dalit Christian, one Dalit Muslim (Pasmanda Muslim), one Ad-Dharmi (Dalit Sikh), one individual from the Other Backward Classes (OBC), and one from a Scheduled Tribe (ST) of central India.

Although we do recognise that caste-based hate speech is now a global problem, for this report, we spoke to respondents primarily based in India, except two, one of whom is based in the USA and other in Austria. They both argued that there are no national boundaries when it comes to caste-based hate speech online.

Given the role played by social media platforms in arbitrating the norms and proliferating of online expression, it is important to explore their commitments towards combating OCHS. For the purposes of our research, we based our selection of platforms based on usage by the interviewees, and those identified by our literature review as critical to understanding OCHS. The platforms for consideration therefore were: Facebook, Twitter, YouTube and Clubhouse. While the former three platforms have had a relatively prolonged presence in the Indian internet market, amassing large user-bases, Clubhouse has been a new entrant. Despite its short period of existence however, Clubhouse has amassed considerable attention for being a hotbed for islamophobic, sexist and casteist hate speech[15], and therefore, warranted a similar level of analysis as Facebook, Twitter and YouTube.

15. Nandy, A. (2021, June 17). *Islamophobia to Casteism, How Hate Thrives Unchecked on Clubhouse.* TheQuint. https://www.thequint.com/news/india/clubhouse-twitter-spaces-hate-speech-islamophobia-casteism-bullying.

By analysing the community standards and other publicly available policy documents of these platforms, we investigated how they address hate speech and abuse directed at individuals on account of their caste. We analysed the platforms in three steps. First, we tried to understand whether the community standards of these platforms explicitly aim to moderate or remove OCHS. We went through the latest version of their community standards in detail, noting whether they mentioned denigrations or abusive posts targeting caste as hate speech. Given that Facebook, Twitter, and YouTube have had more than 8 years of presence in India, we also sought to understand the timeline of their policy evolution with respect to caste, i.e., when did they decide to list caste as a protected category?

Second, we examined whether they operationalised their community standards for caste protection via user interfaces (UI) that allowed affected individuals to report harmful and abusive speech as OCHS. Third, we analysed platform transparency reports, wherever available, to understand the overall enforcement of community standards related to hate speech. More specifically, we sought to see if the enforcement information available for hate speech standards was as granular as the community standards themselves.

Online Caste-Based Hate Speech

In India, hate speech against marginalised caste groups has been legitimised by religious texts, social customs, rituals and cultural practices.[16] Notions of 'purity of "upper-caste" groups, and conversely of 'pollution' of "lower-caste" groups have made the latter subject to discrimination, violence, and dehumanisation.[17] On the other hand, "upper-caste" networks maintain their caste privilege – of political, social, and economic advantage – by structurally excluding lower caste members.[18]

Everyday language has further normalised casteist slurs to shame and criminalise Dalit bodies, occupations, and even their attire. Examples include caste-based slurs used in proverbs that ridicule Dalit women for wearing slippers or sandals.[19] Other examples include insults for assigning their children 'proper' names,[20] and everyday language that associates crimes, such as thefts or robbery, with specific castes, whose names are then used as insults. In still other cases, Dalits are called 'cunning jackals' – a slur that claims that they cannot be trusted if they are of 'fair' complexion.[21]

In the context of Hindu nationalism, the existing offline network is translated to the online medium through its elites who are also a growing part of the diaspora. While optimistic accounts regarding the internet focus on the aspect of the 'voice' of marginalised communities, it is also necessary to look at the 'gaze' of the state-corporate-dominant caste nexus while analysing the politics that play out in these online spaces.[22] Through this lens, dominant groups can be seen as using social media as a tool to counter Dalit assertion.[23] As Tejas Harad points out, many popular social media accounts "extol the virtues of upper-caste practices, cement the Hindu-Muslim binary, deny the reality of the caste system and dish out a propagandist historical account of India's past."[24]

Dalit assertion is no stranger to this sort of backlash, which mostly takes the form of violent attacks and ridicule.[25] While physical attacks are not possible on the internet – though we know of exchanges online that have led to attacks offline[26] – ridicule and abuse act as online proxies of physical altercations.

16. Kumar, V. (2005). Situating Dalits in Indian Sociology. *Sociological Bulletin*, 54(3), 514–532
17. Berg, 2020. Op. cit.
18. Vaghela et al., 2020. Op.cit
19. Ibid.
20. Traditionally, Dalits could not give their children names that reflected the naming conventions of the "upper castes", and were forced to assign them garbled names, such as Gattu Ram, Seva Ram, or Takku. The proverb "Bitiya chamaar ka naam Rajraniya" is a exclamation of surprise that a Dalit (chamaar) named their daughter 'Rajrani' (an "upper-caste" name). See Kumar, 2005, Op.cit.
21. Ibid
22. Therwath, I. (2012). Cyber-hindutva: Hindu nationalism, the diaspora and the Web. *Social Science Information,* 51(4), 551–577. https://doi.org/10.1177/0539018412456782.
23. Harad, T. (2018, August 31). *Towards* An Internet Of Equals. Mint. https://www.livemint.com/Leisure/c7XqIj7NcWEhmcV3Wdeaul/Towards-an-internet-of-equals.html.
24. Ibid
25. Berg, 2020.Op. cit.
26. One of our interview respondents, a professor in Delhi University, told us about two instances where he escaped being attacked by right wing student affiliates in response to his caste-related posts online.

The online sphere has also led to the emergence of covert harassment and abuse: casteists use terms like 'reservation people' or 'quota fellows' to refer to Dalits to indicate that they do not deserve to be as educated or socially mobile as they are.[27]

In its 2019 report, Equality Labs also reached similar conclusions: 40% of all casteist hate speech they encountered on Facebook was regarding reservation.[28] This highlights a 'double stigma' effect, where lower castes are stigmatised not only for their caste identities but also for availing of constitutionally-mandated affirmative action.[29]

While incendiary speeches have resulted in the actual eruption of violence and genocides targeting specific communities, it is also important to look at hate speech not only as an incitement to violence and a threat to public order, but also as a violation of one's right to dignity.[30] Hate speech and abuse have the potential to leave deep psychological wounds, which further lead to feelings of self-hatred and humiliation in victims. Hence, the capacity of words to hurt can be physical. Hate speech is often inflicted with the intention of harm.[31]

Until we consider the violation of dignity a significant part of caste-hate speech, a large part of online hate speech will remain under-studied and unmoderated.

Taking a broader view of such speech-acts and accomodating for incitement to discrimination and attacks on dignity, the International Dalit Solidarity Network (IDSN) defined caste-hate speech as "any communication form such as speech, writing, behaviours, codes, signs, or memes that manifest hierarchies, invoke humiliation, serve to dehumanise, incite discrimination, degrade self-worth or perpetuate discrimination and are often the sources of physical, mental or material violence to a person or a group based on caste identity."[32]

27. Pilot interviews for the project, July 2021.
28. Soundararajan, 2019. Op.cit.
29. Deshpande, A. 2015. *From formal to substantive equality.* Seminar 672.
30. Sajlan, 2021.Op.cit.
31. Matsuda, M. J., Lawrence, C. R., Delgado, R., & Crenshaw, K. W. (1993). *Words That Wound: Critical Race Theory, Assaultive Speech, And The First Amendment*. Routledge & CRC Press.
32. Shanmugavelan, Dr. M. (2021). *Caste-Hate Speech: Addressing Hate Speech Based* On Work And Descent. International Dalit Solidarity Network (IDSN). https://idsn.org/wp-content/uploads/2021/03/Caste-hate-speech-report-IDSN-2021.pdf.

Caste hate speech in domestic and international law

India has several legal provisions that criminalise certain forms of caste-hate speech and abuse.[33]

Passed in 1989, and amended in 2015**eduled Castes and Scheduled Tribes (Prevention of Atrocities) Act** criminalises several discriminatory and hateful acts against members of Scheduled Castes and Tribes (SC/STs). The Act criminalises any expression that promotes hatred or ill-will towards SC/STs[34] or disrespects any deceased persons they hold in high regard.[35] Apart from acts that target the communities at large, they also criminalise acts against individuals. These include imposing or threatening a social or economic boycott[36] and insulting, intimidating, or abusing a person with their caste name to humiliate a person "within public view".[37]
The consensus on the meaning of 'public view' across several courts is that the place is within public view even if it is a private space:[38] all that matters is that the utterances/acts are heard/observed by others.[39] Thus, most provisions of the SC/ST Atrocities Act also apply to the online space.[40]
The **Protection of Civil Rights Act** was enacted in 1955, in line with constitutional objectives, to abolish untouchability. The Act prohibits expressions that incite or encourage any person or group to practise untouchability in any form, or insults a member of a SC on the grounds of untouchability.[41] A provision in the Act clarifies that such acts shall be presumed to have been committed on the ground of untouchability unless proven otherwise.

The **Indian Penal Code** also contains several provisions relating to caste-hate speech and abuse. Section 153A criminalises promoting "disharmony or feelings of enmity, hatred or ill-will between [...] castes", and committing any act that causes fear or alarm to any caste. Section 153B criminalises expression that asserts that a particular caste cannot "bear true faith" to India or promotes the view that a particular caste group should be denied their rights. Notably, criminal defamation also specifically includes a reference to caste: expression intended to harm the reputation of a person in relation to their caste is considered defamation.[42] Note that these provisions have been subject to widespread criticism given their misuse to target critics of religion and political dissenters.[43]

---

33. Arun, Chimayi, et al (2018). *Hate Speech Laws* In India. Centre for Communication Governance, National Law University, Delhi. https://ccgnludelhi.wordpress.com/2018/05/04/launching-our-mapping-report-on-hate-speech-laws-in-india/ ; Chaudhary, Digvijay (2021). *Survey of Caste-Related Hate Speech Laws in South Asia.* Centre for Internet and Society (forthcoming).
34. Section 3(u), Scheduled Castes and the Scheduled Tribes (Prevention of Atrocities) Act, 1989. Note that this would require a threat to public order. See Sajlan, 2021.Op.cit
35. Section 3(v), Scheduled Castes and the Scheduled Tribes (Prevention of Atrocities) Act, 1989.
36. Section 3(z)(c), Scheduled Castes and the Scheduled Tribes (Prevention of Atrocities) Act, 1989.
37. Section 3(r) and 3(s), Scheduled Castes and the Scheduled Tribes (Prevention of Atrocities) Act, 1989
38. Arun, 2018.Op.cit.; M.A. Kuttappan v. E. Krishnan Nayanar And Another, Criminal Appeal 2192 Of 1996 (High Court of Kerala 21 February 1997). Asmathunnisa vs State Of A.P & Anr, Criminal Appeal 766 of 2011 (Supreme Court of India 29 March 2011). Swaran Singh & Ors vs State Tr. Standing Council & Anr, Criminal Appeal 1287 of 2008 (Supreme Court of India 18 August, 2008)
39. Yunus Daud Bhura vs State Of Maharashtra, 2002 BomCR Cri (Bombay High Court 2 May 2001)
40. In a key decision, the Delhi High Court stated that an online post insulting or intimidating a member of SC/STs would be punishable even if the post's privacy settings were not 'public'. See Ms. Gayatri Apurna Singh
41. This Act is much narrower than the SC/ST Atrocities Act. According to the Act, merely referring to someone's caste (regardless of whether it was with hateful intent) is not considered a crime. See also M.A. Kuttappan vs. E. Krishnan Nayanar and Others Criminal Appeal No. 450 of 1997; Sarita Shyam Dake vs. Sr. Police Inspector, Writ Petition No. 1746 of 2004.
42. Section 499, Indian Penal Code.
43. Bajoria, Jayshree and Lakhdhir, Limda (2016). *Stifling Dissent: The Criminalization of Peaceful Expression in India.* Human Rights Watch. https://www.hrw.org/report/2016/05/25/stifling-dissent/criminalization-peaceful-expression-india

Internationally, the International Covenant on Civil and Political Rights (ICCPR) – which India has ratified – recognises the right to freedom of expression and requires states to legally prohibit "advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence.[44]" The United Nations General Assembly adopted the International Convention on the Elimination of All Forms of Racial Discrimination (ICERD) in 1965,[45] which India ratified as well. The Convention commits its members to eliminating discrimination, including on the basis of race, colour, descent, nationality, or ethnic origin. In 2002, the UN Committee on the Elimination of Racial Discrimination (CERD) asked states to "take measures against any dissemination of ideas of caste superiority."[46] While the UN Strategy and Plan of Action on Hate Speech (2019) did not mention caste as a specific characteristic,[47] the CERD has clarified that caste discrimination is in the scope of the Convention as it is essentially discrimination based on 'descent'.[48]

It is noteworthy that despite domestic constitutional and legal provisions to combat caste discrimination, India has adopted a stance based on a narrow interpretation of the ICERD: it claims that caste discrimination is not in its scope.[49] A change in this stance could pave the way for domestic legislation to align with international standards. Specifically, amendments to the SC/ST Act would be necessary to prohibit expression that dehumanises or sows hatred against specific castes (based on ideas of racial or caste superiority), even if it does not pose a threat to public order.[50]

44. Article 20, International Covenant on Civil and Political Rights. https://www.ohchr.org/en/professionalinterest/pages/ccpr.aspx
45. International Convention on the Elimination of All Forms of Racial Discrimination https://www.ohchr.org/EN/ProfessionalInterest/Pages/CERD.aspx
46. Shanmugavelan, 2021.Op.cit.
47. United Nations Strategy and Plan of Action on Hate Speech (2019). https://www.un.org/en/genocideprevention/documents/UN%20Strategy%20and%20Plan%20of%20Action%20on%20Hate%20Speech%2018%20June%20SYNOPSIS.pdf
48. Sajlan, 2021. Op.cit.
49. Ibid
50. Sajlan, 2021. Op.cit.

OCHS: Perspectives from interviews

This section summarizes and discusses our interview respondents' experiences and insights on caste-hate speech on social media platforms.

Forms and types of caste-hate speech online

Our respondents reported a wide prevalence of casteist abuses and slurs (including rape threats and gendered and queerphobic hate speech towards women and people belonging to queer people). One way to understand OCHS is as a continuation of the offline hate that the respondents encountered growing up; they reported facing similar kinds of slurs, abuses, and derogatory references to their caste identity both offline and online.

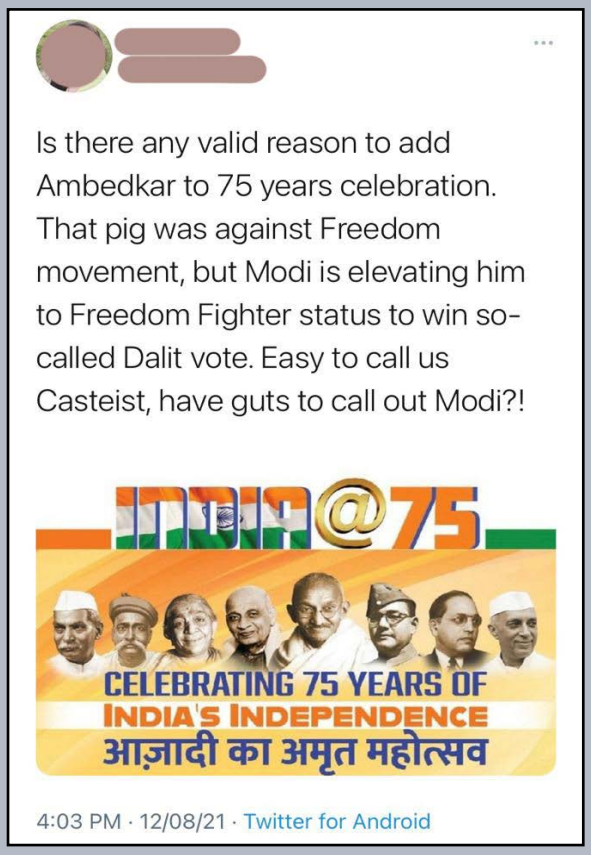
However, such speech manifests differently online. One of the respondents noted that offline, people would not usually comment on appearance – they would not call other people ugly or dark; online, however, they would. Another respondent said that unlike physical spaces, where individuals from marginalized castes were organised and resilient to 'upper-caste' hate, social media made individual targeting easier and harder to counter.

One of the most common terms used in casteist abuse online are 'quotawala' and 'reservation-wala' (one who avails constitutionally mandated quotas and reservations in public institutions).

The phrases are intended to undermine a lower-caste person's worth by insinuating that they are somehow less deserving of education or employment because their admission into an institution came with relaxations. One of our respondents said that this could be very damaging to children and teens on social media as it can contribute to a self-perception that they are lacking in comparison to their peers.

Source: Twitter
***Figure 1:** Dalits being mocked and shamed for availing the constitutional right of reservation. (**Translation:** What logic! Such long propaganda... It has only one meaning: Dalits are equal to reservation. Problem solved!)*

Insults to Dr. B.R Ambedkar were also very common; they were used to undermine those who were politically vocal and fought for Dalit rights. One of our respondents said that just having the name 'Ambedkar' in his username on Twitter attracted a lot of hate from people who blamed Ambedkar for being divisive, being against India's fight for independence, and responsible for spreading caste hatred in India. There were memes on Ambedkar where another respondent was specifically tagged in order to harass and rile him. Another respondent, also a prolific worker for Dalit rights, said that they received overwhelming hatred for their celebration of Ambedkar Jayanti[51], an occasion for which Twitter introduced a special emoji. They said that the hate hurt and shocked them. Ambedkar was not just a Dalit leader, but an intellectual and drafter of the Constitution, they said; he should be above such trolling because of his identity.



Source: Twitter
***Figure 2**: Hate Speech against Dr. B.R.Ambedkar*

51. Ambedkar Jayanti is widely celebrated in India and globally on 14th April – Dr. B. R. Ambedkar's birthday – to commemorate his achievements, activism, and scholarly work

Another form of OCHS manifests when abusers twist inconspicuous words into slurs. Again, these can only be identified by the person they target. For example, one of our respondents said: ***Some slang words are continuously used... you won't find [them] in either English or Hindi dictionaries. Sometimes they mix two to three words and create a new one for your caste (for hidden insults). These words get so generalised but we know why they are being used and what they mean... However, if you see it from the perspective of language, then you won't be able to understand from which language they originate because they are twisted to escape from any legal punishment. Some words are not even considered abuse. For instance, we usually see in Rajasthan – an OBC community, Mali, is mostly called Kandha – which is the local word for onion – and if you say Kandha, the other person will understand that you are insulting them.***

Source: Twitter
Figure 3: *Some abusers twist words to humiliate Dalits in a way that it cannot be detected as hate speech; the wordplay on the word 'Dalit' (written here as 'Dull-it') may make it unrecognisable to automated tools used by platforms for filtering OCHS.*

Words like 'Bhimte' – a twist on the greeting 'Jai Bhim', which DBA communities use to celebrate Dr. Ambedkar – are also used as slang to abuse and mock those working for the rights of the marginalised. One respondent told us that his caste name is the word for 'dirt' in his language, and how it is used to insult members of his caste.

In the same vein, we found out that overt caste slurs are often made covert to insult an individual's marginalized identity. 'Pasmanda', for instance, is a term that denotes a political organisation that represents Dalit Muslims. A respondent noted that they are often called 'Pasanda', which is a food dish. They said, "***Thus, it becomes difficult to identify whether the person is referring to the caste or the food dish. This may sound [like a] technical difference only, but the intent is clear.***"

Our respondents highlighted these subtler aspects of casteist expression, which make it difficult for DBA communities to express themselves fully online. Social media platforms do not recognise these covert attacks, making any possible remedies difficult.

One of our respondents told us that she is working with a prominent social media platform to make them aware not only of overt casteist slurs in Indian languages, but also of these covert caste- and community-based references used to insult and abuse DBA communities. She said that it is important for platforms to know the contexts in which certain phrases are used as the context determines whether an expression is caste-hate speech.

Silencing and exclusion

The interviews revealed how caste-hate speech is not confined to slurs or abuses alone. As one respondent said, the platform as a whole silences a voice or dismisses it altogether. He said that for him, online caste-hate speech "*... would be [an utterance intended] to dismiss my arguments without taking [them] into consideration or giving any thought to them ... not taking me seriously and silencing me.*"[52] Another respondent concurred; as an openly Dalit Christian female, she finds that all that people can talk about is her Dalit Christian identity and not the content of her argument. Murali Shanmugavelan has discussed this kind of speech as 'outing'.[53] He explains 'outing' as loaded conversations that subtly discriminate against Dalits to make them feel that they are not welcome in a space.[54]

Source: Twitter
Figure 4*: A Tweet targeting a Dalit journalist for talking about the intersections of caste and gender-based violence in the Hathras Rape Case, 2020*
(Translation: You news media people are ignorant. You are idiots who always address the issues of Dalits. You can provide news even without writing about them. This nation is being destroyed by news channels and journalists.)

This silencing, dismissal, and outright abuse are the result of the assumption that social media spaces are elite spaces – that they belong to the urban, English-speaking populace, which very often intersects with the upper-caste population in India.[55] A respondent reported that people often ask him to "go back to the jungle", as he belongs to a tribal community from central India.

One of our respondents also said that Dalits and other marginalised populations can be years late to any social media space where the majority presence is of people from upper-caste communities. The dearth of DBA individuals online results in a lack of resistance to hate against them or the political figures who fought for their rights. This is gradually changing.

In response to being excluded from participating in the elite community, marginalised populations can form their own public spheres, or 'counter publics'.[56] However, on social media, where the tussle between the privileged and oppressed classes is for access to the same space, the former responds to the latter with hate speech. One respondent used the Hindi word 'pratikriya' (reaction) to define the reaction of the upper castes to the entry of marginalised communities to these elite spaces.

52. Interview, July 2021
53. https://www.youtube.com/watch?v=Wni0US_LAr0&ab_channel=AmbedkarKingStudyCircleUSA.
54. Ibid.
55. Vaghela, P., Mothilal. KR., & Pal. J (2020): Indian Political Twitter and Caste Discrimination -- How Representation Does Not Equal Inclusion in Lok Sabha Networks.
56. N. Fraser, (1990). Rethinking the Public Sphere: A Contribution to the Critique of Actually Existing Democracy. *Social Text, 25/26*, 56-80. https://doi.org/10.2307/466240.

Hate as a reaction to assertion of identity

While conducting interviews, we worked with the hypothesis that caste-hate speech is directly related to caste assertion by members of DBA groups online. This is in line with caste-based assertions in the physical sphere, where violent crimes against Dalits escalated with their increased assertion for rights and representation.[57] Our respondents felt the same; one of them summed it up thus:
The reason why we [have] to face such incidents is that we are raising our voices against the so-called established society and in favour of [the] Constitution [of India]. We come from a community that was not even allowed to speak some years back. If you are from [an] upper caste and [you] fight for us, then you will be called the 'Champion of Secularism', but if you are an SC/ST/OBC, then you will be termed casteist [or[unworthy of reservation, or your merit [will] be neglected and then the way people treat you will get harsher.

He added that in the fight for rights for DBA populations, the actual people belonging to these communities are always left behind while it is the upper-caste saviours who get all the attention.

We gathered from the interviews that people with strong, assertive profiles on the internet, who spoke openly about rights-based caste discourse, routinely attracted hateful speech in the form of slurs or abuses on social media. Multiple interview respondents had people tell them to go back where they came from – which essentially meant that they should be performing their caste occupations rather than being online.

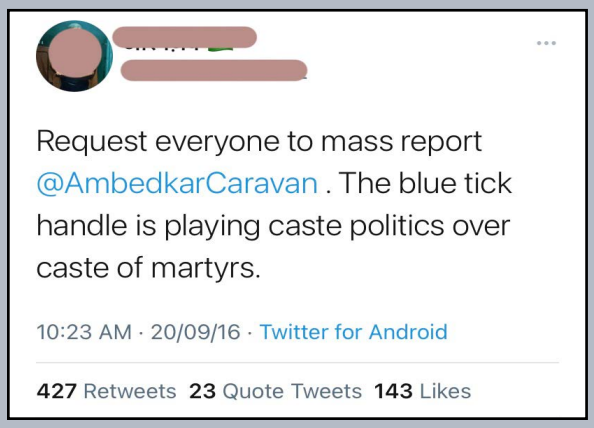
Another respondent said, "hate speech comes to Dalits for merely existing." Many of our respondents shared that simply declaring their marginalised identity online was a surefire way to invite hate. She said that because Dalits, Adivasis, and those who identify as Bahujan are only now present on the internet in such large numbers, the incidence of hate speech is greater now than ever before. The ability to create anonymous user handles, she claimed, has helped create a culture of impunity among those spreading hate.

57. Berg, 2020. Op.cit.

Challenges in the moderation of hate speech

While social media platforms provide options for the redressal of hate speech, they fall short in substance and implementation. The most common reason for this, as most of our respondents pointed out, was that they found gender- or race-based hate speech easier to report than caste-hate. This lack of recognition of caste within the redressal process discourages them from reporting OCHS. We discuss this in greater detail in the next section.

Respondents also said that social media platforms hardly took any action when they reported OCHS. Most of them did not bother checking the outcome of the redressal process, since they were not expecting any results. They mentioned that Twitter sometimes took action against OCHS, but that their process had immense scope for improvement. This is a continuation of their experience of the offline world, where hate speech or crimes are largely overlooked by the police and government. One of our respondents, an influential person on Twitter with a sizable follower count, called up the then director of policy at Twitter to report a post as the standard reporting mechanism did not work for them. The post in question was then taken down. This kind of swift action is only possible if one is a notable personality on Twitter, with access to company personnel. It is not a generic recourse. The respondent said, "***Yes, if you have a blue tick, then your report will be heard; otherwise, they won't even recognise your complaint. This is a hierarchy in itself. Yes, you can surely report but they won't care about it.***"
One respondent brought up a striking reason for not reporting casteist speech on social media: the trend of reverse criminalisation. In the physical world, it is very common for people from DBA communities to be implicated in their own complaints. "What if I am somehow made responsible for the crime?" seems to be a common line of thought among our respondents; it made them refrain from reporting the hate directed at them.



Source: Pradeep Attri
***Figure 5**: A Tweet calling on people to mass report the account of an anti-caste activist in order to silence him from speaking.*

Most social media platforms used by our respondents are headquartered in the US and do not fully understand the cultural and socio-political context of India. Their Indian employees and contractors – including content moderators – themselves tend to come from upper-caste communities and have their own politics, biases, and blind spots while making decisions on the platform. At best, most of them are oblivious to caste discrimination (a privilege of only the upper-castes in India), and at worst, they hold believe that discrimination based on caste is not problematic. This lack of diversity in their teams impacts the moderation process and policies, thereby impacting the contours of what constitutes acceptable speech.[58]

58. Soundararajan, 2019. Op.cit.

The psychological impact of OCHS

Our respondents agreed that caste-hate speech and abuse can take a toll on the mental health of people belonging to DBA communities. Caste discrimination, including hate speech, has been normalised to such an extent that affected communities do not attribute their mental agony to its daily occurrences in their lives. OCHS, especially in digital spaces, where people go in search of close-knit communities, is a reiteration of the dehumanisation that DBA people face every day. Our respondents regularly used the phrases "attack on dignity", "making us feel like animals", "messing with mental health" to explain what they went through when they encountered hate speech. One told us that the stretched mental health fabric of Dalits reaches a breaking point with caste-based slurs and abuses online.

While the respondents said that they do not compromise on what they had to say, they almost always think twice before posting on social media platforms, especially when their content relates to caste or caste-based assertion. They said that even when they tried to respond rationally to hate messages and comments, they found it difficult to engage with abuse.

A number of them have either shut down their social media profiles or stopped posting anything because it is difficult for them to process the hate that comes their way. Some of them also worried about the rise of right-wing Hindu nationalism in India, feared backlash, and thought it best to not post on social media.

A qualitative exploration of the experience of affected communities told us about the extent to which caste-hate speech impacts people everyday. While the number of people from caste-affected backgrounds[59] on the internet has increased over the years, it still makes an inconsequential dent in the systematic culture of abuse online. Apart from studying caste-hate speech and finding solutions to end it, it is important to understand the subjective experiences of those who are targeted. To create any form of redressal, it is critical to identify the heinousness of this phenomenon and the different ways it impacts people on the internet.

59. Shanmugavelan, 2021. Op.cit.

The role of private platforms

Given the extent of our respondent's experiences with OCHS on popular social media platforms like Facebook and Twitter, we now investigate the steps that these platforms have taken to mitigate harm. As we indicated in our methodology, the platforms we have picked for our research are Facebook, YouTube, Twitter and Clubhouse. With the sole exception of Clubhouse, the other three social media platforms have been mainstays in the Indian internet ecosystem for a large part of this decade, with Facebook, YouTube, and Twitter launching their Indian operations in 2006[60], 2008[61], and 2013[62], respectively.

Yet, at the time of their launch as well as at the time of their entry into the Indian market, none of these platforms were equipped with suitable tools to address speech and content issues. This meant that the rules that should have guided the contours of acceptable speech on platforms – or the processes by which platforms ought to have dealt with unacceptable speech – were missing. Examples of this can be found in YouTube's tryst with the Thai government in 2006 and the Turkish government in 2007, where content that had violated the domestic laws in each of these countries was allowed to remain on the platform for prolonged periods.

In both cases, redressal was also on a case-by-case basis, and not in a systematic fashion.[63]
As Kate Klonick has argued, Facebook, Twitter, and YouTube's transition from being software companies to content platforms has been slow; accordingly, the formulation of their speech standards succeeded their entries into different regional markets, and not the other way around.[64] Even when these platforms appointed teams and lawyers to consolidate and centralise their moderation rules, they were noticeably America-centric.[65]

As the Thai and Turkish examples show, for a considerable amount of time, these platforms operated in regional markets, but their moderation rules did not reflect the cultural contexts of the market. Scholars argue that coupled with their tendency to prioritize extremist, viral content for profit,[66] this context-agnostic model for moderation of speech provided the baseline condition for the proliferation of hate speech on these platforms.

60. Sushovan Sircar. (2020, April 25). A 14-Year 'Timeline': Facebook's Roller-Coaster India Journey. *The Quint*. https://www.thequint.com/tech-and-auto/facebook-india-journey-mark-zuckerberg-whatsapp-reliance-jio-rollercoaster-ride#read-more.
61. The Economic Times. (2008). YouTube launched in India. *The Economic Times*. https://economictimes.indiatimes.com/tech/internet/youtube-launched-in-india/articleshow/3017907.cms?from=mdr.
62. Mitter, S. (2015). How Twitter Changed Its Mind On India | Forbes India. *Forbes India*. https://www.forbesindia.com/article/big-bet/how-twitter-changed-its-mind-on-india/39391/1.; 2013 was only the year when Twitter India was formally incorporated. Indians did have access to Twitter before 2013, with Twitter's first Indian user joining the platform in 2006, see here: Mane, U. (2012, 19 July). India's First Twitter User – An Interview With Naina Redhu. *Social Samosa*. https://www.socialsamosa.com/2012/07/indias-first-twitter-user-an-interview-with-naina-redhu/
63. J. Rosen, .(2008, 28 November). Google's Gatekeepers. *The New York Times Magazines*. https://www.nytimes.com/2008/11/30/magazine/30google-t.html
64. Klonick, K. (2018). The New Governors: The People, Rules, and Processes Governing Online Speech. *Harvard Law Review, 131(6)*. https://harvardlawreview.org/2018/04/the-new-governors-the-people-rules-and-processes-governing-online-speech/.
65. Ibid.
66. Khan, L & Pozen, D. (2019). A Skeptical View of Information Fiduciaries. *Harvard Law Review*, 497. https://harvardlawreview.org/wp-content/uploads/2019/12/497-541_Online.pdf

Facebook

Today, Facebook is one of the most popular social media platforms in the country. With over 340 million Indian users on Facebook, India forms its largest market.[67] Despite its 15 years of existence in the Indian market, it was only in 2018 that 'caste' first made an appearance in its community standards about hate speech.[68] In its policy rationale, Facebook explains that direct attacks on people based on 'protected characteristics', which includes race and gender, would be construed as hate speech. Caste was included as part of these protected characteristics on 31 August 2018, as per the change log available on Facebook's website.

On 16 December 2019, Facebook added a list of designated comparisons, generalisations, and behavioural statements that would be considered dehumanising. These included dehumanising comparisons made against, for instance, women and transgender persons. Casteist dehumanisations were not initially part of this list; they were added on 23 September 2020 and then removed from the policy for unspecified reasons on 12 October 2020, only to be added again on 28 January 2021. As of 17 August 2021, as per its existing policy, comparing Dalits or people belonging to SCs or lower-castes to menial labourers is construed to be the most egregious form of hate speech.

Facebook's updated hate speech policy is operationalized via the user-interface (UI) made available to users for reporting problematic content. On selecting the content to be categorised as 'hate speech' [see Figure 1], Facebook allows users to choose denigration of social castes as a ground for classifying content as hate speech [see Figure 2]. While there is no change-log publicly available for ascertaining when this feature was added, its existence lends credence to the notion that Facebook would have information about the amount of content on its platform that relates to casteist hate speech.

General information about the enforcement of Facebook's hate speech policy is found in its transparency reports, which provides data on how the platform enforces its community standards, including hate speech. This includes, for instance, data on how many content restriction requests Facebook received from the government; the number of requests on which it took action; and the number of accounts that were restricted. While the policy on hate speech aspires to be granular on the different ways in which speech can be dehumanising to people with protected characteristics, the same granularity is absent in the information on the enforcement of community standards.[69]

67. https://www.statista.com/statistics/268136/top-15-countries-based-on-number-of-facebook-users/
68. https://transparency.fb.com/policies/community-standards/hate-speech/
69. https://transparency.fb.com/data/community-standards-enforcement/hate-speech/facebook#PREVALENCE

That is, the information about enforcement of 'hate speech' standards does not aspire to break down the information provided on the lines of, say, how many Facebook posts were removed on account of being OCHS, how many of them were appealed, and how many were restored. Given that there is previously documented evidence of Facebook restoring explicit hate speech on their platforms after removal,[70] this absence of granularity obscures important information about their actual rate of removal of OCHS.[71]

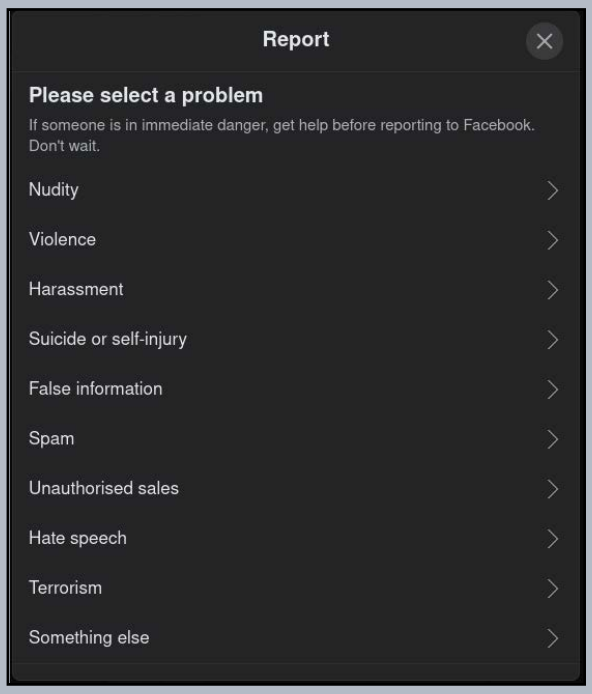


Figure 6: Facebook's UI for reporting content

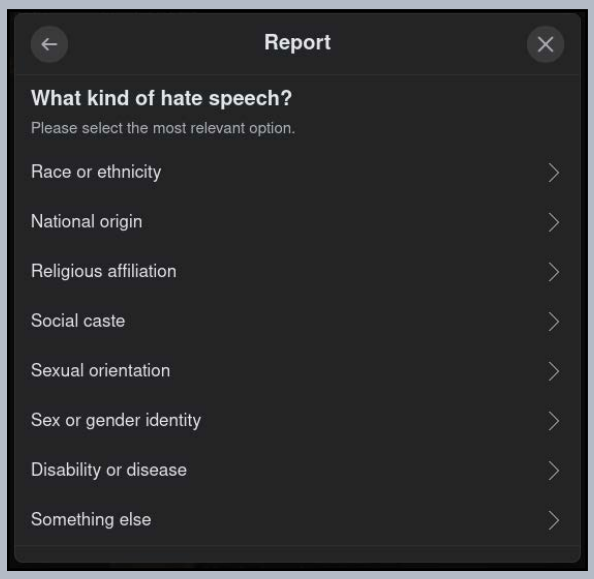
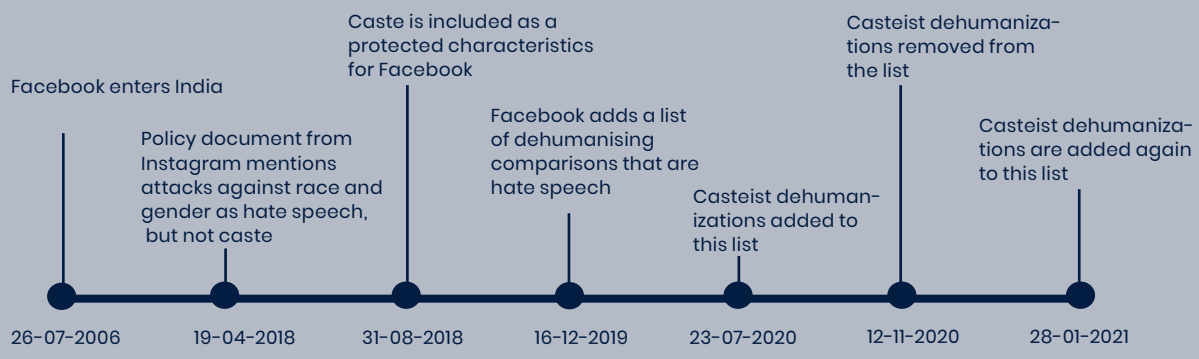


Figure 7: Facebook's UI for selecting hate-speech categories

Facebook also owns Instagram, the photo-sharing social media platform. India is also the largest market for Instagram, with 180 million users as of July 2021.[72] Unlike Facebook, however, Instagram's community standards are not completely centralised and conflicting accounts of their speech norms are available online. For instance, a document from 19 April 2018 mentions that Instagram prohibits hate speech based on certain characteristics, which include race and gender but not caste.[73]



70. Soundararajan, 2019. Op.cit.
71. Ibid.
72. https://www.statista.com/statistics/578364/countries-with-most-instagram-users/#:~:text=As%20of%20July%202021%2C%20India,audience%20of%2093%20million%20users.
73. https://about.instagram.com/blog/announcements/instagram-community-guidelines-faqs

On the other hand, on Facebook's transparency report centre, hate speech is defined with reference to both Facebook and Instagram, and includes a current list of protected characteristics that mentions caste.[74] Instagram's UI for reporting problematic content, however, does not provide a similar list of categories or classifications of hate speech for the user to choose from. Instead, the option to report content as hate speech is couched in broad, general terms [see Figure 8]. Finally, similar to information about the hate speech policy enforcement on Facebook, there is no granularity in the way Instagram approaches different types of hate speech on its platform

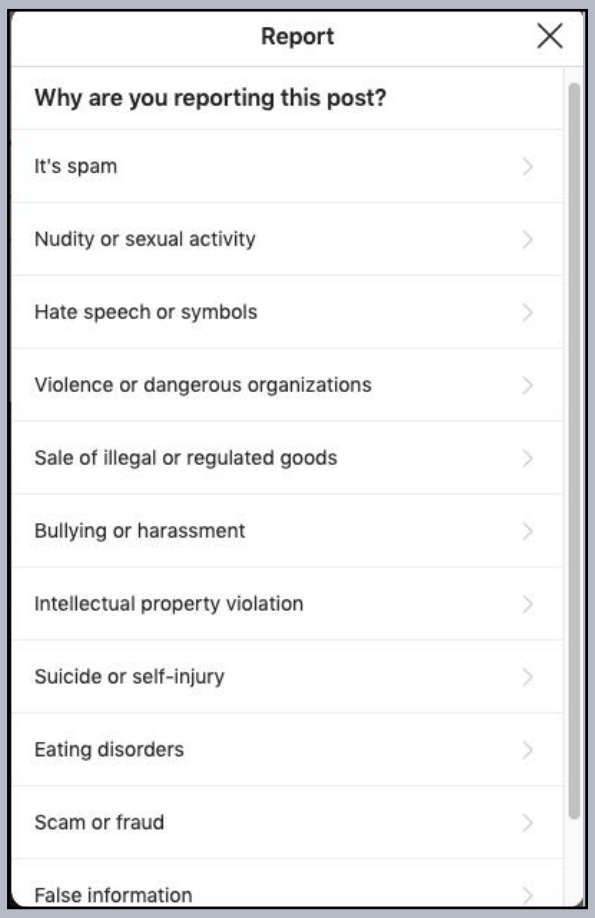


Figure 8: UI for reporting content on Instagram

A note on Whatsapp

A few respondents told us about how hate speech for their posts on social media (Facebook and Twitter) took the form of personal messages on these platforms. There were also mentions of cross-messaging, where hate triggered by a post on Twitter reached their personal messages on Instagram, indicating how people scoped the whole internet to convey their anger in the widest manner possible.

However, one female respondent said that in parts of rural Gujarat, upper-caste men record videos of themselves raping lower-caste women, accompanied with vile casteist abuses, and circulate them on Whatsapp. The rape plus the abuses become doubly traumatic and insulting for the women. Here, hate speech on the internet is not just confined to social media and does not just exclude certain communities from the online space; rather, it excludes them from physical spaces as well. So, in a way, these communities do not just experience hate speech online but suffer its ramifications offline as well.

74. https://transparency.fb.com/data/community-standards-enforcement/hate-speech/instagram/#restored-content

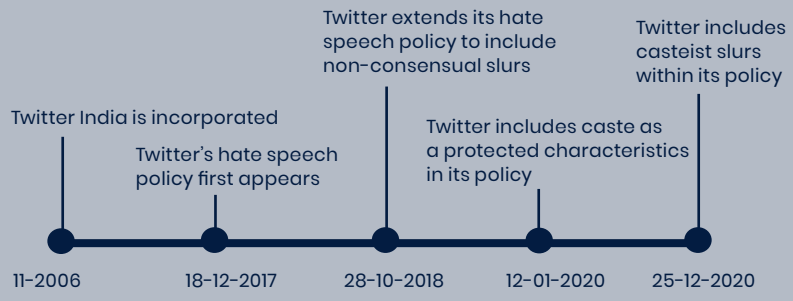
Twitter

As of July 2021, India ranked third in the number of Twitter users worldwide, with about 22 million users.[75] As per the information available on Wayback Machine, a historical archive of web pages, Twitter's hate speech policy first made an appearance on the archive on 18 December 2017. Similar to Facebook's protected characteristics list, this policy prohibits attacks against people based on attributes like race or gender. Caste was not a part of these attributes and was only added on the policy update dated 12 January 2020.[76]

On 28 October 2018, Twitter expanded its policy to include the use of repeated/non-consensual slurs that tend to dehumanise, degrade, and/or establish negative stereotypes about a protected category of attributes.[77] Caste was absent from this list of protected categories for two years; it was added only on 25 December 2020.[78] As of 18 August 2021, 'caste' is now a protected category within Twitter's Hateful Conduct Policy,[79] and calling for segregation, incitement, or dehumanisation against a group of people or individuals based on a protected category is prohibited.

On 18 December 2020, Twitter introduced 'Spaces', a feature allowing users to host and participate in live audio conversations.[80] Given that the format in which content is produced and distributed over Spaces is different from Tweets and Messages, it is not clear how Twitter's community guidelines protecting individuals would apply to content shared over this new medium. The reporting mechanism for Spaces puts the onus of deciding the contours of unacceptable speech on the host or co-host. Only they have the power to mute, block, or remove users from the Spaces they are hosting.[81]

Similar to Facebook, when it comes to information about enforcement of these community standards,[82] Twitter reports the number of accounts reported in violation of its hateful conduct policy, but does not delineate how much action was taken in response to OCHS.



75. https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/
76. https://web.archive.org/web/20200112053811/https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy; The immediately preceding update was on 9 January 2020, and did not mention 'caste', so we can assume that the 12 January 2020 update was the one to introduce it. https://web.archive.org/web/20200109183745/https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy
77. https://web.archive.org/web/20181028023901/https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy
78. https://web.archive.org/web/20201225015808/https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy
79. https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy
80. https://twitter.com/TwitterSpaces/status/1339639767089238019
81. https://help.twitter.com/en/using-twitter/spaces-hosting#reporting
82. https://transparency.twitter.com/en/reports/rules-enforcement.html#2020-jul-dec

Finally, although 'caste' is now an explicit part of Twitter's policy against hateful conduct, users are not allowed to report any posts as potentially casteist. Within Twitter's UI for reporting problematic posts, while there is a category for reporting hate directed against a protected category, caste is visibly absent; instead, it carries the outdated list of categories [see Figure 9]. As a result, any potential progress Twitter has made by recognising casteist dehumanisation as hate speech has become redundant.

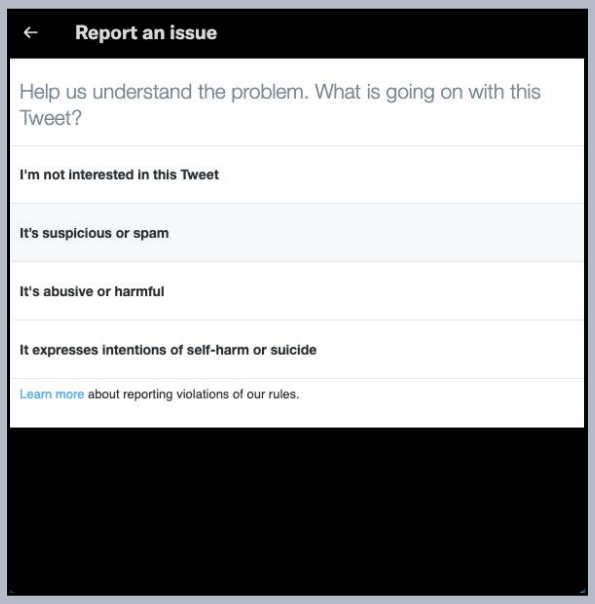


Figure 9: Twitter's UI for reporting content

Youtube

As of July 2021, India topped the list of countries for the most number of YouTube users, with 225 million active users every hour.[83] As per the information available on Wayback Machine, the URL for YouTube's hate speech policy first made an appearance on the archive on 14 November 2019.[84] It prohibits content promoting violence or hatred against any individuals or groups of individuals based on a list of attributes. Caste was a part of this list of protected attributes on this version of the policy and continues to be part of the policy as of 26 August 2021.[85]

YouTube's UI for reporting content is divided into the UI for reporting a video [see Figure 10] and for reporting a comment [see Figure 11]. In both of these, while hateful or abusive content forms a part of the reporting mechanism, there is no mention of the protected attributes from the main policy, thus making it potentially difficult for users to ascertain whether YouTube would deem the problematic content hateful.

On the other hand, data about community standards enforcements provides no additional information about how YouTube deals with hate speech targeted at different protected characteristics.[86]

83. https://www.omnicoreagency.com/youtube-statistics/
84. https://web.archive.org/web/20191114002846/https://support.google.com/youtube/answer/2801939?hl=en&ref_topic=9282436
85. https://support.google.com/youtube/answer/2801939?hl=en&ref_topic=9282436
86. https://transparencyreport.google.com/youtube-policy/removals?hl=en

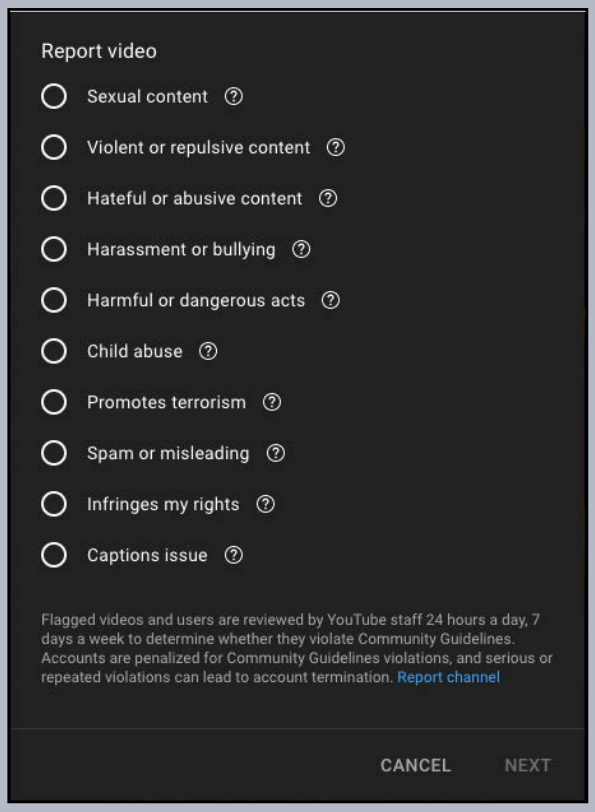


Figure 10: YouTube's UI for reporting a video

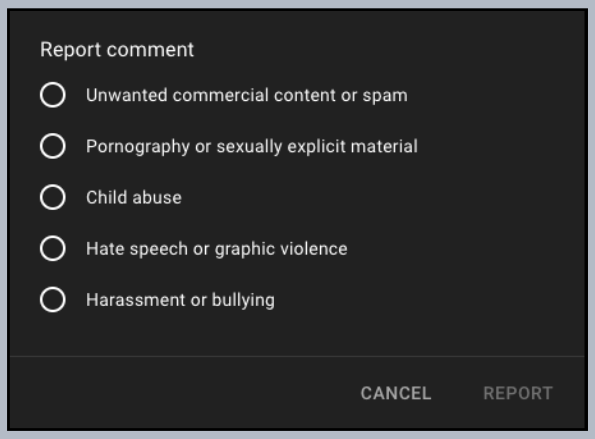


Figure 11: YouTube's UI for reporting a comment

Clubhouse

Unlike the other three platforms we studied, Clubhouse is a relatively new entrant into the Indian internet ecosystem, having been around for only about a year. Further, unlike the other three platforms, whose content is predominantly multi-media, including text, photos, and videos, Clubhouse is a predominantly audio-based platform. This difference in form sets up this platform – and potentially any other platforms with similar content models (including Twitter Spaces) – for a different kind of challenge when it comes to dealing with OCHS, since engagement with content on Clubhouse is ephemeral and real-time, as opposed to content on platforms like Facebook and Twitter, which leaves a more permanent record after publication. The audio in Clubhouse rooms cannot be paused, rewound, or listened to after the broadcast is over.[87] This means that traditional forms of content moderation, including both ***ex-ante*** and ***ex post facto*** moderation, cannot be applied to the audio within these rooms.

Within the short period of its existence, Clubhouse has generated a considerably large Indian user base, with Indian users accounting for about 80% of its downloads between June 1 and 22.[88] On the other hand, it has also drawn criticism for being a hotbed for Islamophobic, casteist and sexist speech,[89] and therefore, its moderation policies warrant a closer look.

87. Aten, J. (2021, 21 February). Clubhouse Is Recording Your Conversations. That's Not Even Its Worst Privacy Problem. *Inc.* https://www.inc.com/jason-aten/clubhouse-is-recording-your-conversations-thats-not-even-its-worst-privacy-problem.html
88. Ananya Bhattacharya. (2021). Clubhouse's next moves in India will determine if it's the next Facebook or the next Foursquare. *Quartz*. https://qz.com/india/2018206/can-clubhouse-last-in-india/.
89. Bose, A. (2021, June 18). *Inside Clubhouse India: Is It The New Ground For Polarisation?* Boom. https://www.boomlive.in/mediabuddhi/clubhouse-india-hate-misinformation-love-jihad-hindutva-muslim-christian-13576; Nandy, A. (2021, June 17). *Islamophobia to Casteism, How Hate Thrives Unchecked on Clubhouse.* *TheQuint*. https://www.thequint.com/news/india/clubhouse-twitter-spaces-hate-speech-islamophobia-casteism-bullying.

Clubhouse's community guidelines[90] state that the platform does not tolerate discrimination based on a host of protected characteristics, including "race, colour, religion".[91] However, 'caste' is not an explicitly protected category within this list. On the other hand, Clubhouse's UI for reporting problematic speech is more detailed than that of both Twitter and YouTube. Figure 12 provides the broad categories within which a user can report the room title, and, interestingly, these categories are far more granular than what is available on their community guidelines page. Further exploration of the hate speech category revealed that Clubhouse allows users to report the title of a room as being casteist [see Figure 13]. We must note, however, that these reporting mechanisms are only restricted to the 'room title', as opposed to the contents discussed within the rooms themselves.

As of the time of authoring this report, Clubhouse did not have a publicly available transparency report. Accordingly, enforcement information regarding these platforms is unavailable.

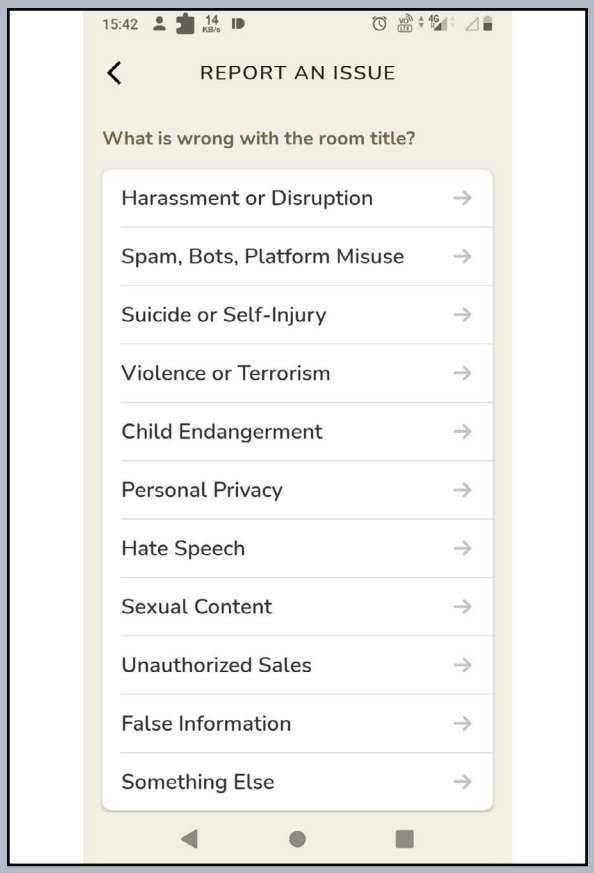


Figure 12: Clubhouse's general UI for reporting problematic speech

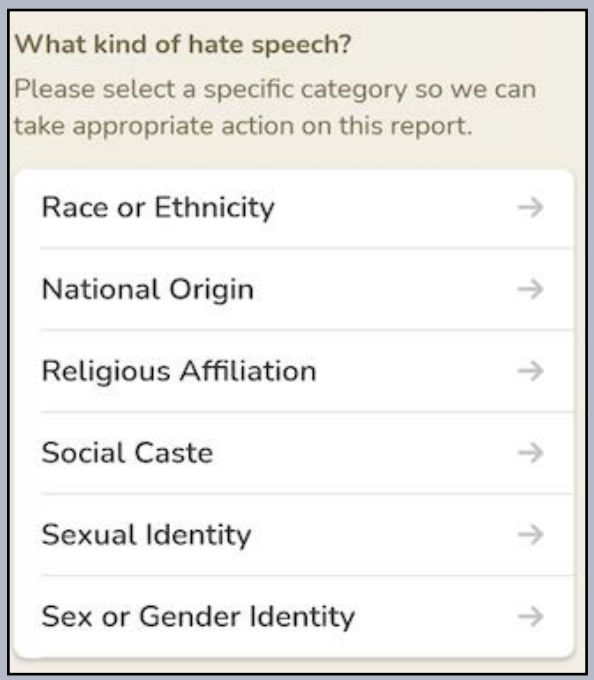


Figure 13: Further categorization of Clubhouse's UI on reporting hate speech

90. https://help.clubhouse.io/hc/en-us/community/posts/360059501212-Community-Guidelines
91. https://community.clubhouse.com/

What can platforms do better?

Our analysis confirms that Facebook, Twitter, and Youtube included caste as a protected category in their hate speech policies many years after they entered India. Even in cases where the platforms committed to prohibiting OCHS on their platforms – as Twitter did –the operationalisation of such commitments has been inadequate. Based on our study of the four platforms, and supplemented by our interviews, we identify two work-streams via which company policies can be overall improved to provide a safer space for their users:

Immediate actions

1. Robust and contextual definitions of hate speech:
In delineating what categories of content are prohibited on account of being hate speech, companies must ensure that these definitions are robust, contextual, and conscious of social conditions in the regions where the platform is functioning. In the context of this report, this means that 'caste' must be made an explicitly protected category, and any forms of denigration or dehumanisation related to caste must be explicitly prohibited from the platform. In case a company owns multiple social media platforms, it must ensure that its standards of speech are uniform across all platforms. And finally, these definitions and lists of protected characteristics must be made accessible to the communities that would make use of them – that is, these standards must be translated into a sufficient number of local languages, among others.

2. Collaboration with DBA and anti-caste activists, academics and organisations:
Platform companies must establish inclusive norms through collaborative efforts in dealing with online caste-hate speech. This necessitates greater representation from communities marginalised based on caste at different levels of the content moderation process. Companies must ensure their participation at all levels, including in policy framing, research, content moderation, data analysis, and technical support. There is an urgent need to engage with communities who are the direct recipients of harmful speech on the incumbent platforms we have studied in this report.[92]

3.The UI should operationalize definitions of OCHS:
A platform's UI for reporting hate speech should reflect a list of protected categories (including caste).

92. *See (2018). Assessing the Human Rights Impact of the Facebook Platform in Sri Lanka,* Article One. https://static1.squarespace.com/static/53bdabe6e4b0b43ac59a9b44/t/5eb97cbe9f56f9201f233649/1589214398996/Sri+Lanka+HRIA_+Executive+Summary_FINAL.pdf; (2018). *Assessing the Human Rights Impact of the Facebook Platform in Indonesia. Article One.* https://static1.squarespace.com/static/53bdabe6e4b0b43ac59a9b44/t/5eb97ca9b2acbe6aa40cbf62/1589214377636/Indonesia+HRIA_+Executive+Summary_FINAL.pdf; (2018). Rebalancing Regulation of Speech: Hyper-Local Content on Global Web-Based Platforms. Berkman Klein Center for Internet and Society Blog. https://medium.com/berkman-klein-center/rebalancing-regulation-of-speech-hyper-local-content-on-global-web-based-platforms-1-386d65d86e32

4.Information about enforcement of community standards:
In enforcing community standards, platforms should aspire to be more granular about how they deal with hate speech produced against different protected characteristics. This may include disaggregating the information on the enforcement of hate speech standards, the lines between different protected characteristics, and the action taken on each category, including appeals and restoration of content.

5.More diversity at the workplace:
Platforms need to hire more DBA personnel so that they can add their analyses, critiques, and/or experiences of hate speech during moderation (both in-person moderation and training algorithms for moderation).

Long-term actions

Beyond these immediate action items, however, much is still left to be done. For instance, the respondents' fears about counter-repercussions on reporting OCHS cannot be addressed through simple commitments on the company's part to combat hate speech, since it reflects both the materialisation of an offline trend into the online sphere and the dangers of malicious coordinated behaviour.

The proliferation of OCHS (and all other forms of hate speech) on these platforms must also be viewed against a tapestry of numerous other instances of speech-related problems on social media platforms, including disinformation and extremism. Each of these instances reflects a fundamental truth about social media platforms: their very nature is a kind of moderation decision that tends to prioritise virality, which in turn produces an endless loop of content-related crises.[93]

These are problems inherent to these platforms; they cannot be solved simply by increasing the list of protected characteristics to encompass caste. Steps towards resolving these problems must necessarily include measures such as online awareness campaigns in which companies highlight their commitment to tackling caste-hate speech on their platforms and the measures they've undertaken to do so. Similarly, companies must be more transparent and accountable with their processes in a more qualitative fashion, beyond reporting the number of posts, accounts, or pieces of content on which action has been taken. This might include information about the net investments they have made towards building inclusive platforms, potential collaborations with organisations working towards similar causes, and an overall openness towards acknowledging the responsibility to ensure that they provide equal and safe spaces to people from marginalised caste groups.

93. Grimmelmann, J. (2018). *The Platform is the Message* (SSRN Scholarly Paper ID 3132758). Social Science Research Network. https://papers.ssrn.com/abstract=3132758.

Conclusion

This report presents insights into online caste-hate speech, combining qualitative accounts of the targets of such expression on social media platforms with an analysis of these platforms' speech norms. Our research shows the wide gulf between the promises made by social media platforms towards providing safe spaces for their users and the implementation of such promises. Especially for large US-based companies, our research also provides insight into the initial disregard for socio-cultural contexts of the markets they operated. We hope these findings provide further guidance for researchers and activists, and actionable steps towards making these platforms relatively safer for people belonging to historically marginalized communities.

There are, of course, clear limitations with the scope of our study. As our methodology indicates, our respondent pool was limited, and accordingly, the entire gamut of India's regional, and cultural variances, in the way people experience casteist hate speech, could not be captured via the interviews. This limited respondent pool also capped our ability to fully explore the intersection of gender, sexuality and caste. In the future, we would be interested in understanding how occupying more than one marginalized identity impacts the experience of respondents, and what possible stakeholder responses could be to these challenges. Finally, while we summarized the legal provisions prohibiting casteist hate speech, we did not extend our research to enquire whether the instances of hate speech used within the report would be considered illegal under these laws. Given that our recommendations are targeted at the social media platforms, this legal analysis is currently out of the report's scope.

The experience and insights of our research participants reveals a snapshot of how caste manifests online, and the harms of casteist hate speech and abuse to marginalised groups. The recent disclosures made by Facebook whistleblowers, and the follow-up investigations into Facebook's processes, compel us to confront the same gamut problems that have been highlighted by this report: severe inaction when it comes to hate speech against minorities, and a lack of cultural context when it comes to moderating content in some regions.[94] While technological solutions and companies cannot solve deep social problems such as caste, this evidence shows that companies have disregarded the concerns of certain publics.[95] Companies' promises of user 'safety', in practice, currently cater to a narrow audience. For a large part of the world, these companies still need to demonstrate, through concrete investments and engagement, their willingness to take action and fight the menace of hate speech.

94. (2021, 28 October). The Facebook Papers and their Fallout. *The New York Times.* https://www.nytimes.com/2021/10/25/business/facebook-papers-takeaways.html
95. *See* Arun, Chinmayi. (2021). *Facecbook's Faces.* (Forthcoming) Harvard Law Review Forum Volume 135, https://ssrn.com/abstract=3805210 or http://dx.doi.org/10.2139/ssrn.3805210

Appendix 1: Indicative questions we asked the respondents

a) How would you understand OCHS?

b) Which social media platforms do you predominantly use?

c) Have you been targeted with OCHS, or other kinds of harassment, violence, or other problematic speech online on any of these platforms?

   i) Have you seen other DBA people around you be targeted with OCHS?
   ii) Do you think being DBA has any relationship with harassment or hate speech online?
   iii) Is your experience across platforms the same or different? In your experience, is there more or less OCHS content on some platforms?
   iv) Also, where do you encounter OCHS most? Do you receive abusive content more in your personal inbox, or do you see it more on public posts?

d) Content:
   i) What kind of OCHS do you face? For instance, is it personal, does it use common caste, class, gender stereotypes, or is it just generally abusive?
   ii) Are you comfortable sharing the details of some of this content?

e) How did you respond to OCHS?
   i) Did you report it on the platform? If yes, under which category? Did the platform respond? How did they respond? Was their response satisfactory?
   ii) Did you report the instance to the police or any other state authority? Did the police/agency respond, and was their response satisfactory?

f) Did you in any way modify your behaviour online to avoid OCHS?

   i) Are you aware of any other DBA social media users modifying their behaviour to deal with OCHS?
   ii) Do you try to create counter-speech?

g) In your opinion, what are some solutions to deal with/respond to OCHS?

   i) What can companies and the government do to address these concerns and make social media safer for users?
   ii) Have you ever seen any platform moderate/hide/mark any word/post that was offensive?
   iii) While dealing with OCHS, did you ever think of something which platforms can incorporate to tackle hate speech – such as automatically removing offensive content, blocking, making offensive content difficult to read, or anything else?