Caste-based hate speech moderation on Facebook (Meta) and Twitter (X)

– Manish Kumar
School of Information, University Of California Berkeley

Keywords: caste; hate-speech; Facebook; Meta; Twitter; X; content moderation; social media; Equality; AI bias; categorization

1. Abstract

The paper will investigate the complexities involved in moderating casteist content on major social media platforms. Focusing on the challenges of categorizing hate speech targeting India's marginalized caste-oppressed minorities, the research aims to delve into the mechanisms utilized by platforms for content moderation. It will critically analyze existing categorization systems, exploring the efficacy and limitations in identifying and addressing casteist comments. By leveraging information sciences concepts like grounded coding and taxonomies, the study will elucidate the complexities of classifying caste-related hate speech within these systems. Moreover, it will highlight the implications of inadequate categorization on content visibility and community impact.

2. Introduction

Unfortunately for the people living in the Indian subcontinent, casteism is still a big issue. It goes back to ancient India and was transformed by various rulers in medieval, early modern and modern eras, especially by Mughal rule and British Raj. The term caste is derived from the Portuguese word casta, meaning "race, lineage, breed" and, originally, "'pure or unmixed (stock or breed)". Originally not an Indian word, it is now widely used in English and Indian languages, closely translated to varna and jati*(Caste System in India, n.d.)*. Caste is very intricately intertwined with Indian society and values. Casteism however abolished by the Indian constitution has still found ways to manifest in our everyday lives. Caste affiliations help determine things like occupation, wealth, network, resources, marriage prospects and respect in the society.

We have seen caste manifest itself in online spaces. A recent study has concluded that thirteen per cent of posts having hate content on Facebook in India pertain to caste-based hate speech, including ‘caste-based slurs, derogatory references to caste-based occupations such as manual scavenging, anti-Ambedkar posts, etc. *(Equality Labs, n.d.)* Shreeti Shubham reveals the existence of private exclusive Brahmin groups which are rife with hate speech. The objective of the private caste groups is to serve only the interests of their caste, which will ultimately reinforce casteism in both the digital and social spheres. *(Shubham, n.d.)* By affiliating themselves with a caste group in the digital sphere, these groups get the most advantage in the offline spheres like housing jobs etc. Many anti-reservation and anti-SC-ST Act (1989) memes are widely shared among the upper caste groups. They call themselves “deserved’ as opposed to the ‘reserved’ category and question the merit of every Dalit. Brahmins called themselves Buddhijeevi (Intellectual), and the structural hatred against seeing the Dalit community in the dominant spaces made them mock Dalits by calling them “Aarakshanjeevi” (A person who survives on reservation), “Bheemta” (an offensive word to demean Dalit Ambedkarites), Chamar, Bhangi (Casteist slurs), and so on.

Effectively moderating and identifying casteist content on social media platforms is an enormous challenge, particularly when it comes to India's caste-oppressed minorities who are marginalised and oppressed. The current methods and techniques used by these platforms to recognise, classify, and mitigate hate speech aimed at these communities frequently fail to take cultural sensitivity into account and are unable to distinguish between subtle forms of discrimination related to caste. Consequently, it is imperative to examine the shortcomings of current content moderation frameworks and suggest workable solutions to improve the identification and responsible handling of casteist content on these platforms.

The scope of this study centres on evaluating the content moderation strategies employed by major social media platforms concerning caste-based hate speech directed at India's marginalised caste-oppressed minorities. It focuses on an in-depth examination of these platforms' categorization and identification mechanisms to discern and mitigate casteist content. The research aims to provide insights into the specific strategies, algorithms, and policies adopted by platforms like Facebook and X formerly Twitter. Additionally, the study seeks to propose recommendations for enhancing content moderation frameworks to address the cultural nuances of caste-related discrimination online. Emphasizing cultural sensitivity, the research endeavours to elucidate the complexities of casteist content moderation within the Indian social context.

Social media networks currently employ a mix of human and artificial intelligence (AI) moderation. We will get into more detail in our Platform analysis. Information sciences research on the moderation of casteist content on social media goes beyond traditional information classification and community management models. This research explores the integration of AI-based content moderation techniques in addition to closely examining the effectiveness of categorization systems. Its goal is to investigate how human judgement and algorithmic decision-making can work together to detect and reduce hate speech related to caste. The study sheds light on the difficulties and developments in reporting of caste-based hate comments on social media.

This study presents a comprehensive investigation into the pervasive issue of caste-based discrimination within the digital sphere, particularly on major social media platforms such as Facebook and Twitter. Beginning with an exploration of the historical roots of casteism and its evolution into modern-day digital spaces, the study critically examines the content moderation strategies employed by these platforms, scrutinizing their effectiveness in identifying, categorizing, and mitigating casteist content. It further outlines a review of community guidelines, practical evaluations of reporting mechanisms, comparative analyses across platforms, and qualitative assessments.

3. Literature review

Facebook now includes “caste” when describing protected classes in hate speech. Facebook India report published by Equality Labs reveals numerous key findings: *(Equality Labs, n.d.)* It studies different reporting workflows out of which only one had “social caste”, that too was not reproducible. Each reporting mechanism offers more or less detail. Which reporting workflow you will see next is unpredictable. The fact that Facebook is testing new workflows on marginalised Indian minority communities is alarming. How can Facebook monitor how well it is protecting these communities on its platform if these categories aren't broken down in the reporting of hate speech and violence? The standards of Facebook India have not changed in line with the company's growth into markets in the Global South. The intricate religious and socio-political contexts in India necessitate a distinct review and co-design process to effectively tackle safety concerns. Crucially, the context is dynamic; the ever-changing memes and stories at the core of hate speech require a similarly dynamic response to contain them.

Barett PM in 2020 *(Barrett, n.d., #)* performed research to find out that Facebook holds the highest number of moderators around 15,000 followed by YouTube with 10,000 moderators and Twitter has around 1500 moderators. They have an average of 30-40 seconds per post. On top of human moderators, AI approaches are also used. However, the present automated system that deploys ML and deep neural networks for the detection and classification of detrimental content has considered accuracy, precision and recall as performance metrics. None of the systems *(U. Gongane et al., n.d.)* to the best of our knowledge have reported the time taken by an algorithm to detect objectionable content. NLP and neural network models show increased accuracy when they are trained to detect a particular type of detrimental content like abusive speech. These models show decreased accuracy when they are applied across different detrimental content formats, languages and contexts. Considering the practical deployment of these algorithms in real-time, time is an inevitable parameter in automated systems.

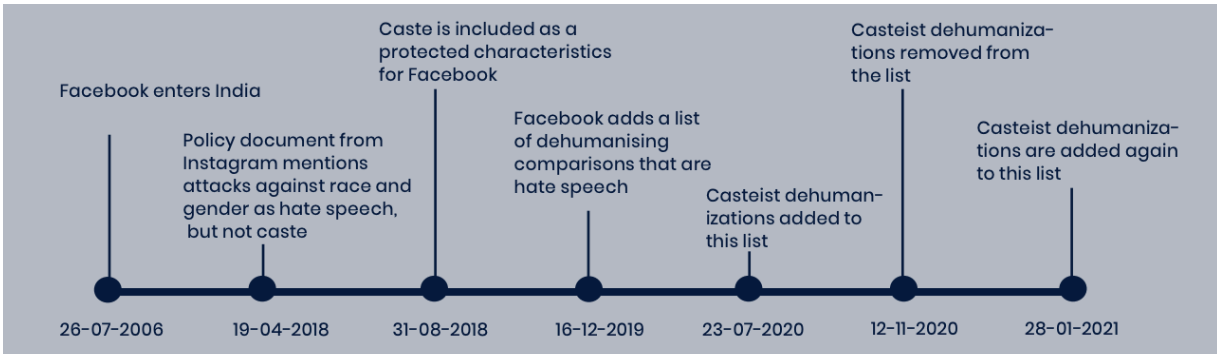
The development of The ComMA Dataset V0.2 (*The ComMA Dataset V0.2 Dataset*, 2021) a multilingual dataset annotated with a fine-grained tagset marking various forms of aggression and contextual elements in conversational threads represents a stride in understanding and combatting hate speech prevalent on social media platforms. This dataset, encompassing 15,000 annotated comments in Meitei, Bangla, Hindi, and Indian English, sourced from diverse platforms including YouTube, Facebook, Twitter, and Telegram, serves as a comprehensive repository for studying aggression and biases manifesting in online discourse. This dataset's paper explains the complex hierarchy of aggression types, which includes biases related to gender, religion, class/caste, and ethnicity/race, all of which are carefully annotated within the context of conversation.

However, despite considerable research exploring the potential of AI-driven solutions for moderating hate speech across various contexts, a noticeable gap exists concerning the specific realm of caste-based hate speech prevalent in India's social media landscape. Notably, while several studies delve into AI-powered content moderation strategies for hate speech, none appear to concentrate explicitly on the nuanced complexities of caste-related discriminatory content within the Indian online sphere. While the work by Equality Labs *(Equality Labs, n.d., #)* has scrutinized caste-based hate speech on Facebook, comprehensive analyses encompassing other prominent social media platforms remain conspicuously absent. The dearth of studies investigating casteist content moderation across diverse social media platforms, apart from the notable exception of the Equality Labs paper focused on Facebook, underscores the pressing need for tailored research efforts aimed at understanding, identifying, and effectively mitigating caste-based hate speech in India's digital sphere.

4. Platform Analysis

4.1 Facebook

Today, Facebook is one of the most popular social media platforms in the country. With over 315 million Indian users on Facebook, India forms its largest market. *(Dixon, 2023)* 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. Caste is now specifically mentioned as a protected category in Facebook's Community Standards, which addresses hate speech based on these traits.


Facebook enters India
Policy document from Instagram mentions attacks against race and gender as hate speech, but not caste
Caste is included as a protected characteristics for Facebook
Facebook adds a list of dehumanising comparisons that are hate speech
Casteist dehumanizations added to this list
Casteist dehumanizations removed from the list
Casteist dehumanizations are added again to this list
26-07-2006
19-04-2018
31-08-2018
16-12-2019
23-07-2020
12-11-2020
28-01-2021

(Kain, n.d.)

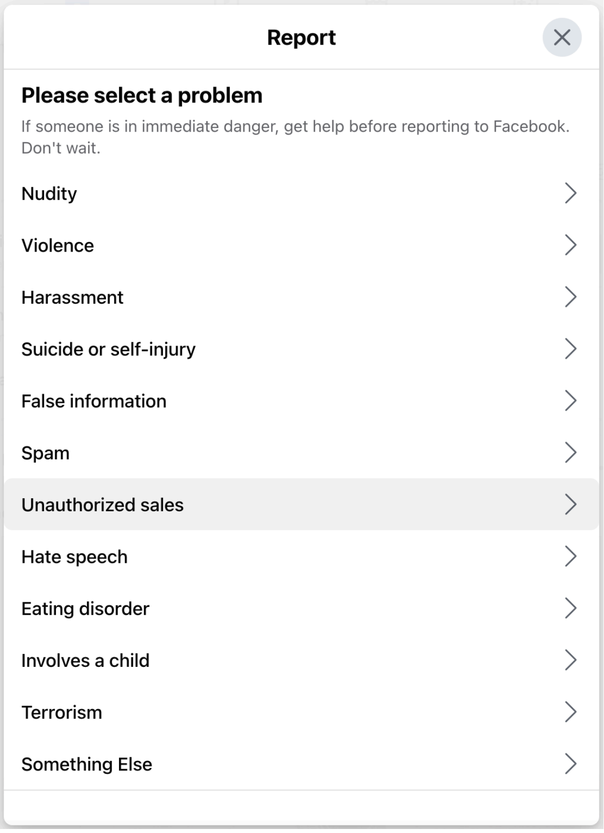
“We define hate speech as a direct attack against people – rather than concepts or institutions – on the basis of what we call protected characteristics: race, ethnicity, national origin, disability, religious affiliation, **caste**, sexual orientation, sex, gender identity and serious disease.” *(Meta, n.d.)*

On their page, they also define how hate speech fares for 3 tier system of Facebook’s guidelines. *(Meta, n.d.)* Taking a closer look one could think of examples of caste-based hate speech based on all definitions be it Tier 1, 2 or 3. Tier 1 is defined as content targeting a person or group of people (including all groups except those who are considered non-protected groups described as having carried out violent crimes or sexual offences or representing less than half of a group) based on their aforementioned protected characteristic(s) or immigration status with insects, animals, filth and harmful stereotypes, etc. They have even mentioned an example of violations that belong to Tier 1: “references to Dalits as menial labourers.” Tier 2 is described as content targeting a person or group of people based on their protected characteristic(s) with Expressions of contempt (in written or visual form), Expressions of disgust, cursing, etc. Tier 3 is Content targeting a person or group of people based on their protected characteristic(s) with segregation and exclusion.

They do acknowledge that defining hate speech is a complex task. It requires understanding the nuances like context and intent. They mention that they are using AI to scale up, but they are a long way from being able to rely on machine learning and AI alone to handle the complexity involved in assessing hate speech.

Facebook’s reporting mechanism:

When you report a post on Facebook, a menu of options appears. You are prompted to select the type of hate speech when you click on Hate Speech. After that, the choice to select Social caste is displayed. Choosing any other hate speech option takes you to the same submission page.

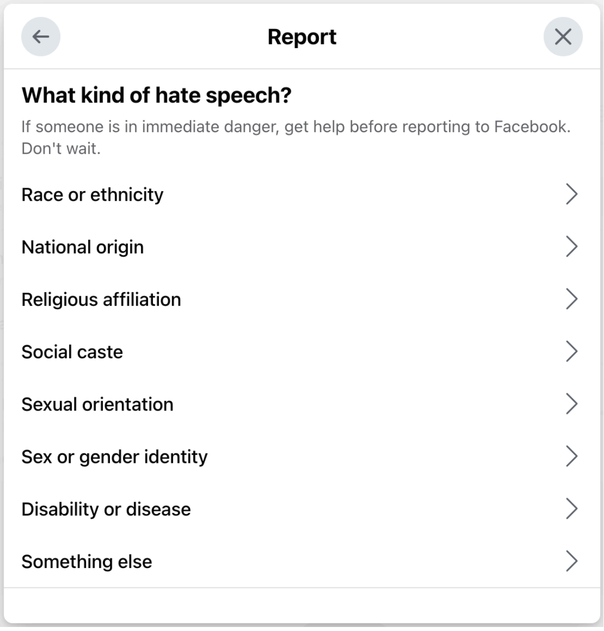


Report

Please select a problem

If someone is in immediate danger, get help before reporting to Facebook. Don't wait.

- Nudity
- Violence
- Harassment
- Suicide or self-injury
- False information
- Spam
- Unauthorized sales
- Hate speech
- Eating disorder
- Involves a child
- Terrorism
- Something Else

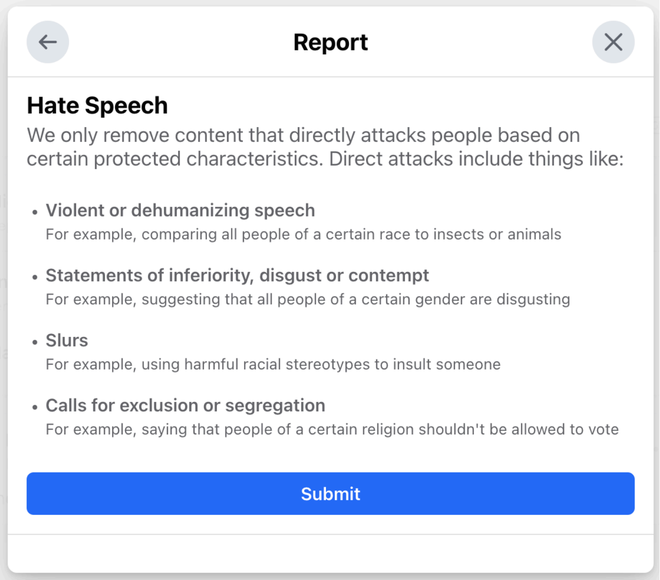


Report

What kind of hate speech?

If someone is in immediate danger, get help before reporting to Facebook. Don't wait.

- Race or ethnicity
- National origin
- Religious affiliation
- Social caste
- Sexual orientation
- Sex or gender identity
- Disability or disease
- Something else



Report

Hate Speech

We only remove content that directly attacks people based on certain protected characteristics. Direct attacks include things like:

- **Violent or dehumanizing speech**
  For example, comparing all people of a certain race to insects or animals
- **Statements of inferiority, disgust or contempt**
  For example, suggesting that all people of a certain gender are disgusting
- **Slurs**
  For example, using harmful racial stereotypes to insult someone
- **Calls for exclusion or segregation**
  For example, saying that people of a certain religion shouldn't be allowed to vote

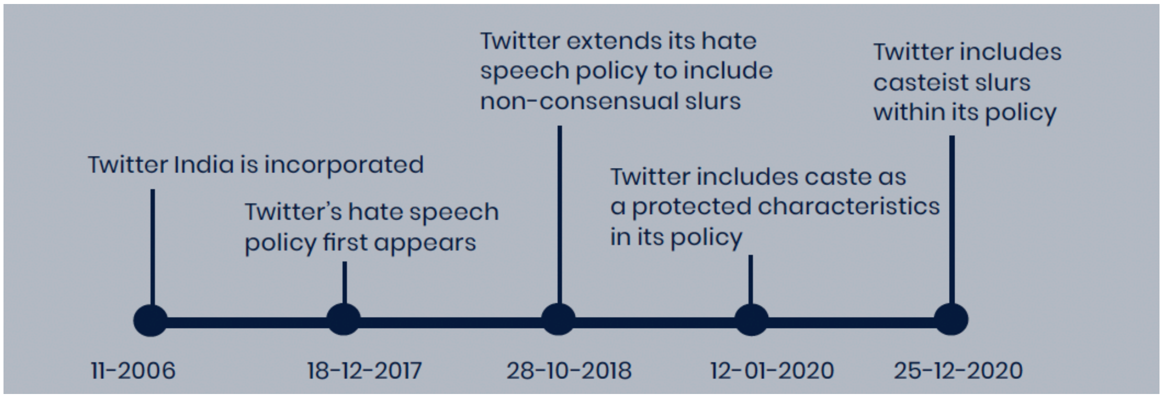
Submit

Examining the Equality Labs report *(Equality Labs, n.d.)*, we find that there was no consistent option on Facebook for reporting hate speech based on caste. This is a significant change. The challenges in accurately identifying and addressing complex forms of hate speech, including those related to caste, through automated means alone are highlighted by Facebook's acknowledgement in their discourse on hate speech recognition using AI. The issue of relying on AI and machine learning to grasp the intricacies of caste-based hate speech could present hurdles given the complexity of context and intent required in assessment, as emphasized in Facebook's discussions on hate speech moderation. Facebook now will be able to track caste and religious hate speech disaggregated by category. Such data is critical to identifying hate speech by scope and scale. Data transparency on this front will allow experts in civil and public society to effectively track and contribute to mitigating hate speech. But this report also mentions that 93% of reported posts on Facebook remain on Facebook, now with all these measures, we must analyse if these have been effective.

Facebook's public claims of removing over 90 per cent of hate speech starkly contrast with internal communications revealing a mere 3 to 5 per cent takedown rate. *(Giansiracusa, 2021)* This discrepancy was unearthed amid revelations from whistleblower Frances Haugen and the subsequent leaks of internal documents. While Facebook parades its proactive hate speech detection rate highlighting how effectively AI detects and removes such content—the real measure that matters, the takedown rate, remains hidden. This deceitfulness emphasizes the limitations of relying solely on AI for content moderation. The leaked documents revealed that more than 95 per cent of hate speech remains on the platform, a staggering statistic that undermines Facebook's touted hate speech moderation effectiveness. Despite claims of improvement, the platform's prevalence metric, signalling a 0.05 per cent of hate speech, fails to represent the true extent of harm in specific communities and user experiences.

4. 2 Twitter

As of January 2023, India ranked third in the number of Twitter users worldwide, with about 27 million users. *(Countries With Most X/Twitter Users 2023, 2023)* 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 to the policy update dated 12 January 2020. *(Web Archives, n.d.)* 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. *(Web Archives, n.d.)* As of 18 August 2021, 'caste' is now a part of Twitter's Hateful Conduct Policy, *(X, n.d.)*. Calling for segregation, incitement, or dehumanisation against a group of people or individuals based on a protected category is prohibited.

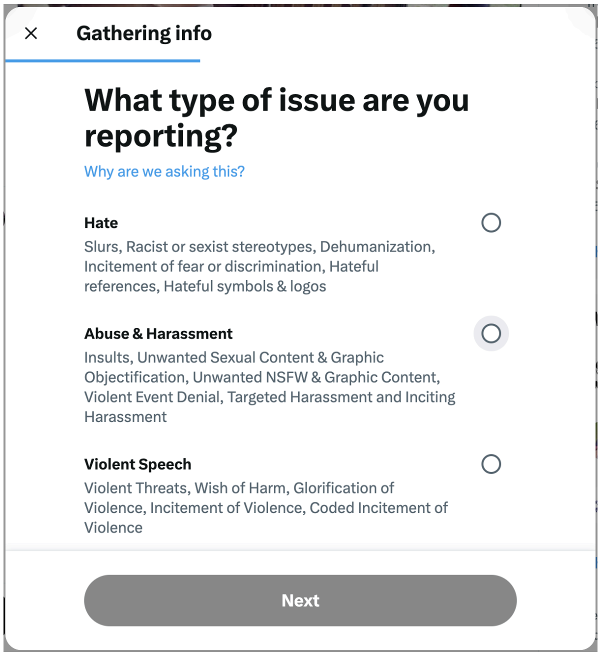

Twitter India is incorporated
11-2006
Twitter's hate speech policy first appears
18-12-2017
Twitter extends its hate speech policy to include non-consensual slurs
28-10-2018
Twitter includes caste as a protected characteristics in its policy
12-01-2020
Twitter includes casteist slurs within its policy
25-12-2020

(Kain, n.d.)

X, previously Twitter said that “X does not tolerate behaviour that promotes violence against, threatens, or harasses other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease.” (X (Twitter)) On its Hateful Conduct page, *(Twitter Help, n.d.)* it describes that it takes action against reports of accounts targeting an individual or a group of people with behaviours such as incitements, slurs and tropes, dehumanisation, hateful imagery, etc. It claims to protect the protected category from these actions in direct messages as well as posts. What precisely falls under the **protected category** is unclear. assuming that it alludes to the protected class as specified by either the Californian or US constitutions. Of which caste is not one protected category.

Twitter reporting mechanism:

Selecting the type of issue is one of the options available to you on the first screen when you report a post. Here, hate speech that is based on caste seems to be the most aligned. Slurs and tropes, dehumanisation, hateful references, hateful imagery, and incitement are among the options listed under hate speech. Caste is only covered by dehumanisation among these options. The remaining choices protect the defined classes and the protected category, but not caste. Therefore, the user is unable to select the precise issue when reporting hate speech based on caste on Twitter. Tweets containing incitement, derogatory images based on caste, and casteist slurs are far too frequent. Users are deterred from reporting incidents as a result, and the platform's ability to gather information about the number of people being affected by a specific type of hate.



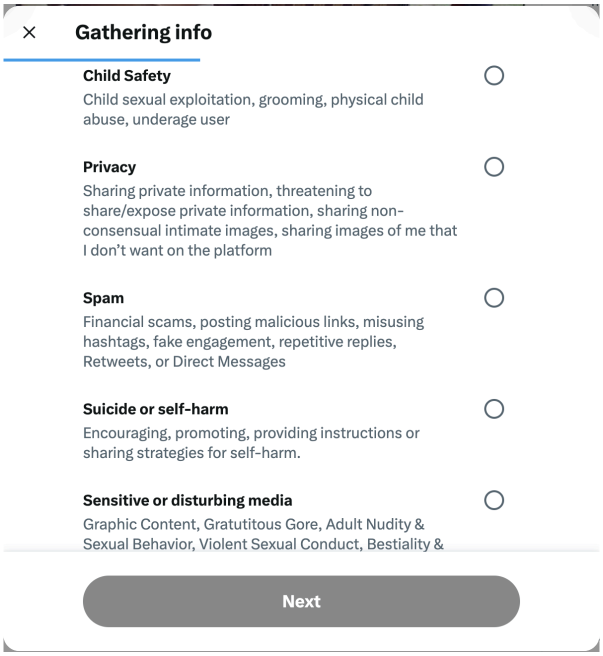
What type of issue are you reporting?

Why are we asking this?

Hate ○
Slurs, Racist or sexist stereotypes, Dehumanization, Incitement of fear or discrimination, Hateful references, Hateful symbols & logos

Abuse & Harassment ○
Insults, Unwanted Sexual Content & Graphic Objectification, Unwanted NSFW & Graphic Content, Violent Event Denial, Targeted Harassment and Inciting Harassment

Violent Speech ○
Violent Threats, Wish of Harm, Glorification of Violence, Incitement of Violence, Coded Incitement of Violence



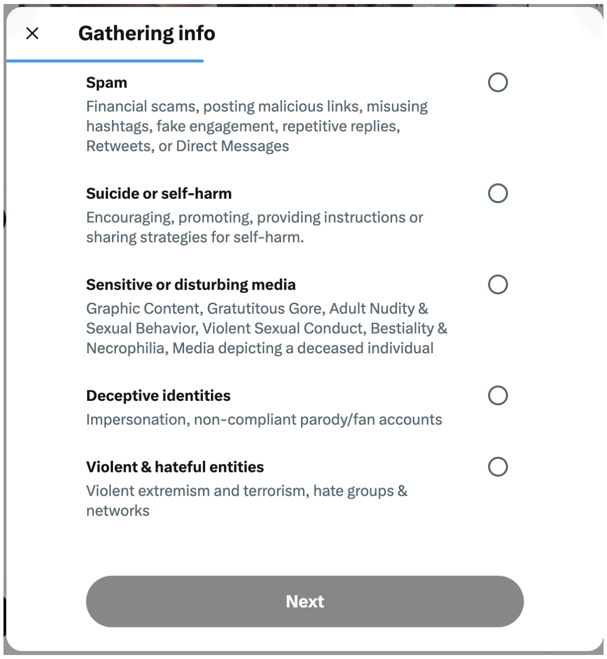
Child Safety ○
Child sexual exploitation, grooming, physical child abuse, underage user

Privacy ○
Sharing private information, threatening to share/expose private information, sharing non-consensual intimate images, sharing images of me that I don't want on the platform

Spam ○
Financial scams, posting malicious links, misusing hashtags, fake engagement, repetitive replies, Retweets, or Direct Messages

Suicide or self-harm ○
Encouraging, promoting, providing instructions or sharing strategies for self-harm.

Sensitive or disturbing media ○
Graphic Content, Gratutitous Gore, Adult Nudity & Sexual Behavior, Violent Sexual Conduct, Bestiality &



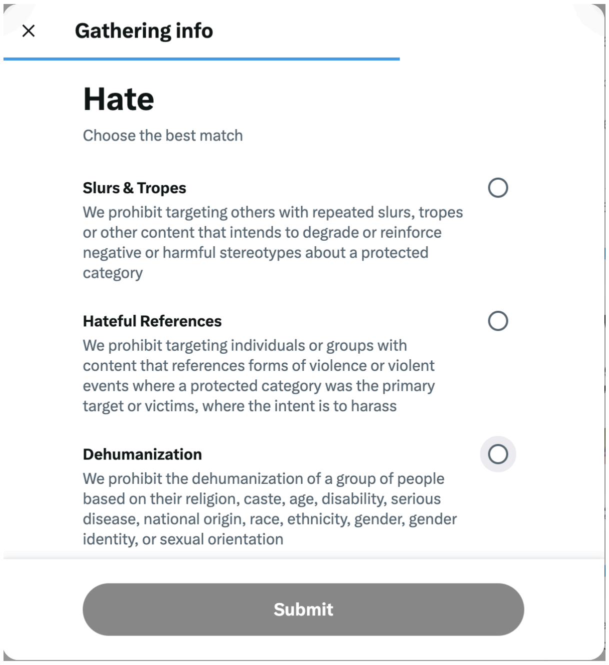
Spam ○
Financial scams, posting malicious links, misusing hashtags, fake engagement, repetitive replies, Retweets, or Direct Messages

Suicide or self-harm ○
Encouraging, promoting, providing instructions or sharing strategies for self-harm.

Sensitive or disturbing media ○
Graphic Content, Gratutitous Gore, Adult Nudity & Sexual Behavior, Violent Sexual Conduct, Bestiality & Necrophilia, Media depicting a deceased individual

Deceptive identities ○
Impersonation, non-compliant parody/fan accounts

Violent & hateful entities ○
Violent extremism and terrorism, hate groups & networks



Gathering info

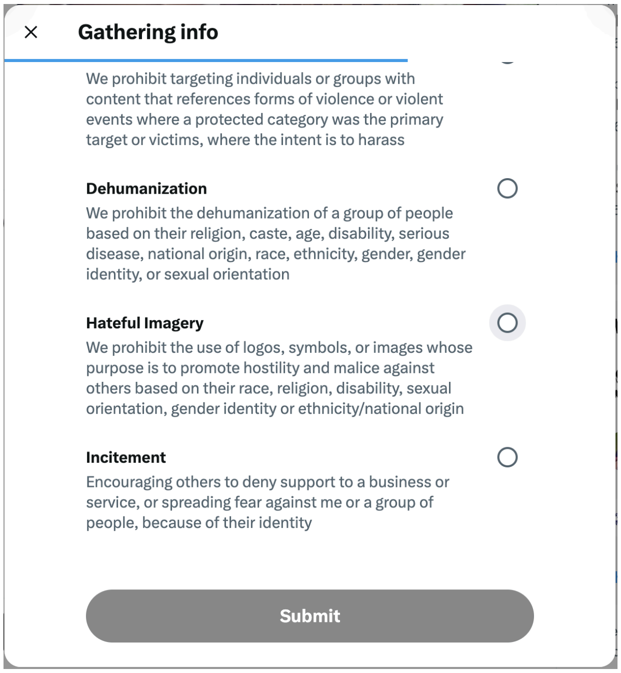
Hate

Choose the best match

Slurs & Tropes ○
We prohibit targeting others with repeated slurs, tropes or other content that intends to degrade or reinforce negative or harmful stereotypes about a protected category

Hateful References ○
We prohibit targeting individuals or groups with content that references forms of violence or violent events where a protected category was the primary target or victims, where the intent is to harass

Dehumanization ○
We prohibit the dehumanization of a group of people based on their religion, caste, age, disability, serious disease, national origin, race, ethnicity, gender, gender identity, or sexual orientation



Gathering info

We prohibit targeting individuals or groups with content that references forms of violence or violent events where a protected category was the primary target or victims, where the intent is to harass

Dehumanization ○
We prohibit the dehumanization of a group of people based on their religion, caste, age, disability, serious disease, national origin, race, ethnicity, gender, gender identity, or sexual orientation

Hateful Imagery ○
We prohibit the use of logos, symbols, or images whose purpose is to promote hostility and malice against others based on their race, religion, disability, sexual orientation, gender identity or ethnicity/national origin

Incitement ○
Encouraging others to deny support to a business or service, or spreading fear against me or a group of people, because of their identity

5. Challenges and Recommendations

There has been a long-running debate on whether caste-based hate speech or casteism should be treated at par with racial-based hate speech or racism. Since caste-based discrimination is covered within the scope of ICERD(International Convention on the Elimination of All Forms of Racial Discrimination) *(Sajlan, n.d.)*, Although the Indian Government Maintains that it is not, caste-based hate speech must be treated with importance. It affects 1/5th of the world's population. Further, to guard against violation of free speech standards, Indian courts can begin interpreting the Atrocities Act in consonance with ICERD. Social media as we saw in the case of Twitter still does not state clearly if caste is a protected category. If these categories are not disaggregated in the reporting of hate speech and violence, how can these social media track its effectiveness in assuring the safety of these communities on its platform? Casteism is no longer an Indian problem, it has become a global issue. So, western countries like the US and the European Union must consider including caste as a protected category. Moreover, there needs to be more transparency concerning the content moderation practices and algorithms used by social media platforms.

In the age of social media when each second so much of content is generated, the platforms rely on automated algorithms to scale hate speech moderation. The use of multimedia forms like images, videos, or GIFs(Graphics Interchange Format) makes this task all the more difficult.

The platforms also continue to use human moderators and that raises another issue related to trauma and exploitation of people living in different geographical locations like the U.S., Philippines, India, Ireland, Portugal, Spain, Germany, Latvia, and Kenya. *(Barrett, n.d.)* The present automated systems are dependent on datasets which are created by annotators which has a potential risk of biased decision by the annotator in assigning a label to content. One should also consider the process of automating the annotation which will add true essence to the complete automation process.

Content Moderation, which should include hiring practices, contractor demographics, and slur lists. These lists should be open and transparent to the public. Empowering an independent audit team that is approved and monitored by both civil society and Internet Freedom advocates as well as by Social Media. This audit team must have clear competencies in caste, religious, and gender/queer minorities and includes members of Indian minorities in its composition. Approaches to addressing error in machine-learning systems, moreover, are fundamentally different from due process protections aimed at ensuring a just result. Machine-learning tools can be evaluated on their “accuracy,” but “accuracy” in this sense typically refers to the rate at which the tool’s evaluation of content matches a human’s evaluation of the same content. This kind of analysis does not address whether the human evaluation of the content is correct “accuracy”, in this context, does not reflect the assessment of ground truth. *(Llansó, n.d.)*

Mark Zuckerberg's repeated emphasis on AI improvements as the panacea for harmful content contradicts the stark reality revealed by internal documents. *(Giansiracusa, 2021)* The company's persistent use of misleading metrics, such as the proactive rate, serves to distract from the concealed takedown rate—a more accurate measure of hate speech removal. As the focus shifts towards addressing the issues of social media, transparency regulations mandating the publication of takedown rates for different categories of harmful content could be pivotal. Such regulations could compel platforms to provide honest and accurate insights into their content moderation practices, preventing deceptive tactics like those exposed within Facebook's hate speech moderation metrics. The need for a more comprehensive approach beyond AI reliance is evident, prompting a call for legal mandates to enforce transparency in content moderation practices across social media platforms. A variety of rationales support the presumption against prior censorship in human rights law, including concerns that: (a) systems of administrative prior censorship bring too much speech into the scope of government review, (b) When censorship becomes convenient it becomes too common, and (c) such systems are too shielded from public scrutiny about what the rules governing speech are and how they are being applied. (Thomas I., n.d.) Therefore we must find a balance between the two.

6. Conclusion

The intricate web of caste-based discrimination within the Indian subcontinent, deeply rooted in historical legacies, continues to pervade modern society, making its presence felt across various facets of life. This pernicious societal ill has seamlessly transitioned into the digital landscape, finding a new platform in social media, where hate speech and derogatory references towards marginalized castes have become all too common. The alarming prevalence of caste-based hate speech on platforms like Facebook and Twitter reveals the systemic challenges in moderating and categorizing such content effectively.

The research critically examined the mechanisms employed by Facebook and Twitter to combat casteist content, shedding light on the limitations and inadequacies within their existing content moderation frameworks. The disparities in categorization and identification mechanisms on platforms like Facebook and Twitter, despite their public commitments to combat hate speech, remain glaring. These platforms' evolving policies, albeit acknowledging caste as a protected category, still lack precision and consistency in addressing nuanced forms of caste-related discrimination, especially within the Indian socio-political context.

Facebook's inclusion of caste within its protected characteristics in hate speech policies signifies a step forward, yet inconsistencies in its reporting mechanisms raise concerns about their efficacy. Despite claims of proactive hate speech detection and removal, internal disclosures unveiled a stark contrast, revealing a mere 3 to 5 per cent takedown rate, raising doubts about the platform's actual effectiveness in curbing hate speech. Twitter, while acknowledging caste within its hate speech policy, faces similar challenges in effectively categorizing and addressing caste-based hate speech due to limitations in reporting options.

The study highlights the urgent need for a nuanced approach to combat caste-based hate speech, emphasizing cultural sensitivity and understanding within content moderation frameworks. Recommendations emphasize the necessity of transparency regulations mandating the publication of takedown rates for various categories of harmful content. Empowering independent audit teams, comprising members from marginalized communities and civil society, can foster more inclusive and accurate content moderation practices.

In conclusion, the struggle against caste-based discrimination in the digital realm demands concerted efforts from social media platforms, policymakers, and civil society. Addressing the inadequacies in content moderation mechanisms, bolstering transparency, and cultivating cultural awareness are pivotal steps toward fostering safer and more inclusive online spaces, and transcending geographical boundaries and cultural divides. The imperative lies not just in recognizing caste as a protected category but in effectively dismantling the systemic barriers perpetuating caste-based discrimination within the online world.

7. References

1. Allan, R. (2017, June 27). *Hard Questions: Who Should Decide What Is Hate Speech in an Online Global Community?* Meta. Retrieved December 12, 2023, from https://about.fb.com/news/2017/06/hard-questions-hate-speech/
2. Barrett, P. (n.d.). Who Moderates the Social Media Giants? A Call to End Outsourcing.
3. *Caste system in India*. (n.d.). Wikipedia. Retrieved December 11, 2023, from https://en.wikipedia.org/wiki/Caste_system_in_India#cite_note-31
4. *The ComMA Dataset v0.2 Dataset*. (2021, November 18). Papers With Code. Retrieved December 12, 2023, from https://paperswithcode.com/dataset/the-comma-dataset-v0-2
5. *Countries with most X/Twitter users 2023*. (2023, September 13). Statista. Retrieved December 12, 2023, from https://www.statista.com/statistics/242606/number-of-active-twitter-users-in-selected-countries/
6. Dixon, S. J. (2023, August 29). *Facebook users by country 2023*. Statista. Retrieved December 12, 2023, from https://www.statista.com/statistics/268136/top-15-countries-based-on-number-of-facebook-users/
7. Equality Labs. (n.d.). Facebook India TOWARDS THE TIPPING POINT OF VIOLENCE: CASTE AND RELIGIOUS HATE SPEECH.

8. Giansiracusa, N. (2021, October 15). *How Facebook Hides How Terrible It Is With Hate Speech*. WIRED. Retrieved December 12, 2023, from https://www.wired.com/story/facebooks-deceptive-math-when-it-comes-to-hate-speech/
9. Giansiracusa, N. (2021, October 15). *How Facebook Hides How Terrible It Is With Hate Speech*. WIRED. Retrieved December 12, 2023, from https://www.wired.com/story/facebooks-deceptive-math-when-it-comes-to-hate-speech/
10. Kain, D. (n.d.). *Online caste-hate speech: Pervasive discrimination and humiliation on social media*. Global Freedom of Expression. Retrieved December 12, 2023, from https://teaching.globalfreedomofexpression.columbia.edu/resources/online-caste-hate-speech-pervasive-discrimination-and-humiliation-social-media
11. Llansó, E. J. (n.d.). *Emma J. Llansó, No amount of "AI" in content moderation will solve filtering's prior-restraint problem*. PhilPapers. Retrieved December 12, 2023, from https://philpapers.org/rec/LLANAO
12. Meta. (n.d.). *Hate Speech*. Transparency Center. Retrieved December 12, 2023, from https://transparency.fb.com/en-gb/policies/community-standards/hate-speech/
13. Sajlan, D. (n.d.). Hate Speech against Dalits on Social Media. https://www.jstor.org/stable/pdf/48643386.pdf?refreqid=fastly-default%3Afa6d739d4cb78b0ca2b39a05fdee3ac9&ab_segments=&origin=&initiator=&acceptTC=1
14. Shubham, S. (n.d.). Caste and the Digital Sphere.
15. Thomas I., E. (n.d.). *The System of Freedom of Expression by Emerson, Thomas I.: Good Plus Paperback (1971) First Thus.* | *Recycled Books & Music*. AbeBooks. Retrieved December 12, 2023, from https://www.abebooks.com/first-edition/System-Freedom-Expression-Emerson-Thomas-I/12583353532/bd

16. Twitter Help. (n.d.). *X's policy on hateful conduct | X Help*. Twitter Help Center. Retrieved December 12, 2023, from https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy

17. U. Gongane, V., V. Munot, M., & D. Anuse, A. (n.d.). Detection and moderation of detrimental content on social media platforms: current status and future directions.

18. Web Archives. (n.d.). Twitter Rules and policies Hateful conduct policy. https://web.archive.org/web/20200112053811/https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy

19. Web Archives. (n.d.). Twitter Rules and policies Hateful conduct policy. Twitter Rules and policies Hateful conduct policy

20. X. (n.d.). *X's policy on hateful conduct | X Help*. Twitter Help Center. Retrieved December 12, 2023, from https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy

21. X (Twitter). (n.d.). *Authorized to represent*. Twitter Help Center. Retrieved December 12, 2023, from https://help.twitter.com/en/forms/safety-and-sensitive-content/hateful-conduct/legal-rep