



A report from

cdt | Research

# Moderating Tamil Content on Social Media

Aliya Bhatia  
Mona Elswah

May 2025



The Center for Democracy & Technology (CDT) is the leading nonpartisan, nonprofit organization fighting to advance civil rights and civil liberties in the digital age. We shape technology policy, governance, and design with a focus on equity and democratic values. Established in 1994, CDT has been a trusted advocate for digital rights since the earliest days of the internet. The organization is headquartered in Washington, D.C., and has a Europe Office in Brussels, Belgium.

---

**ALIYA BHATIA**  
**MONA ELSWAH**

Authors

# Moderating Tamil Content on Social Media

## CDT Research Report

**Aliya Bhatia**  
**Mona Elswah**

### WITH CONTRIBUTIONS BY

Samir Jain, Michal Luria, DeVan L. Hankerson, Dhanaraj Thakur, Varun Rao, and Kate Ruane.

### ACKNOWLEDGMENTS

We would like to first thank the incredible team at the Centre for Internet and Society. Thank you to our Advisory Committee made up of Afef Abrougui, Juan Carlos Lara, Nanjira Sambuli, and Jillian York. We also thank Tarunima Prabhakar and Aatman Vaidya. All views in this report are those of CDT.

A special thank you to Osheen Siva for our beautiful cover and design work and Tim Hoagland for directing the creative for this project.

We thank the study participants who generously shared their time and insights with us. All quotes have been lightly edited for readability.

This work was made possible through a grant from the Internet Society Foundation.

**Suggested Citation:** Bhatia, A. & Elswah, M. (2025). Moderating Tamil Content on Social Media. Center for Democracy & Technology. [<https://cdt.org/insights/moderating-tamil-content-on-social-media/>]

References in this report include original links as well as links archived and shortened by the Perma.cc service. The Perma.cc links also contain information on the date of retrieval and archive.

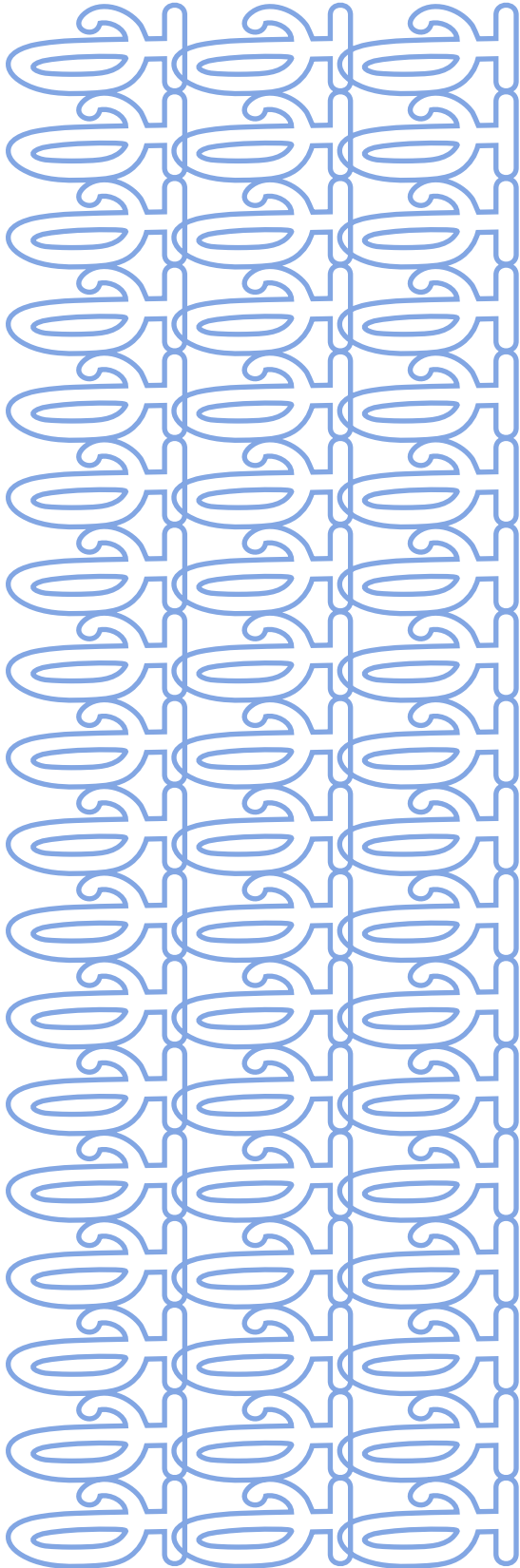


This report is licensed under a [Creative Commons Attribution-ShareAlike 4.0 International Licence](https://creativecommons.org/licenses/by-sa/4.0/)

# Contents

<b>Introduction</b>	<b>5</b>
<b>Tamil as a Low-Resource Language</b>	<b>7</b>
<b>Tamil as a Historically Politicized Language</b>	<b>8</b>
<b>Main Findings</b>	<b>10</b>
1. Tamil speakers use a mix of social media services and online forums in code-mixed, computer-mediated, and transliterated Tamil, sometimes to circumvent moderation.	10
2. Globalized vs. localized approaches: Social media services and online forums pursue a mix of approaches to moderating content in Tamil.	13
3. Many users perceive moderation as an effort to “silence” their voices, sometimes on political grounds.	19
4. Despite advancements in Tamil NLP research and content analysis capabilities for Tamil, social media companies are slow in adopting automated moderation tools for Tamil — due in part to low financial investment and lack of engagement with Tamil NLP experts.	25
<b>Recommendations</b>	<b>29</b>
<b>Appendix</b>	<b>34</b>
Methods and Data Collection	34
<b>References</b>	<b>35</b>

# Introduction



**T**amil is a language with a long history. Spoken by over 80 million people worldwide, or over 1% of the world's population, early inscriptions in the language date back to the 5th Century B.C.E (Murugan & Visalakshi, 2024). The language is spoken widely in India (predominantly in Tamil Nadu and Puducherry), in Sri Lanka, and across diaspora communities in Malaysia, Thailand, Canada, the United Kingdom, the United States, and beyond. Despite the widespread use of the language, there remains limited understanding of how major social media platforms moderate content in Tamil. This report examines the online experiences of Tamil users and explores the challenges of applying consistent content moderation processes for this language.

This report is part of a series that examines content moderation within low-resource and indigenous languages in the Global South. Low-resource languages are languages in which sufficient high-quality data is not available to train models, making it difficult to develop robust content moderation systems, particularly automated systems (Nicholas & Bhatia, 2023). In previous case studies conducted in the series, we found that this lack of high-quality and native datasets impeded effective and accurate moderation of Maghrebi Arabic and Kiswahili content (Elsawah, 2024a; Elswah, 2024b). Inconsistent and inaccurate content moderation results in lower trust among users in the Global South, and limits their ability to express themselves freely and access information.

This report dives into Tamil speakers' experiences on the web, particularly on popular social media platforms and online forums run by Western and Indian companies. We highlight the impact of Tamil speakers' perception of poor content moderation, particularly against a backdrop of democratic backsliding and growing repression of speech and civic participation in India and Sri Lanka (Vesteinsson, 2024; Nadaradjane, 2022). Ultimately, what emerges in this case study is a fragmented information environment where Tamil speakers perceive over-moderation while simultaneously encountering under-moderated feeds full of hate speech.

We used a mixed-method approach, which included an online survey of 147 frequent social media users in India and Sri Lanka; 17 in-depth interviews with content moderators, content creators, platforms' Trust & Safety representatives, and digital rights advocates; and a roundtable discussion with Tamil machine learning and data experts. The methods are detailed in the appendix. Based on these methods, we found that:

1. Tamil speakers use a range of Western-based social media platforms and Indian platforms. Our survey indicates that Western social

media platforms are more popular among Tamil speakers, while local TikTok alternatives are gaining popularity due to India's TikTok ban. Online, Tamil speakers use tactics to circumvent content moderation, employing **“algospeak” or computer-mediated communication, and, at other times, code-mixed and transliterated Tamil using Latin script for ease and convenience.** These tactics complicate moderation.

**2. Tech companies pursue various approaches to moderate Tamil content online, but mostly adhere to global or localized approaches.** The global approach employs the same policies for all users worldwide, and relies on moderators and policy members who are not hired based on linguistic or regional expertise. Moderators are assigned content from across the world. In contrast, the local approach tailors some policies to meet Tamil language-specific guidance, and relies on more Tamil speakers to moderate content. Some Indian companies employ a hybrid approach, often making occasional localized adjustments for Tamil speakers.

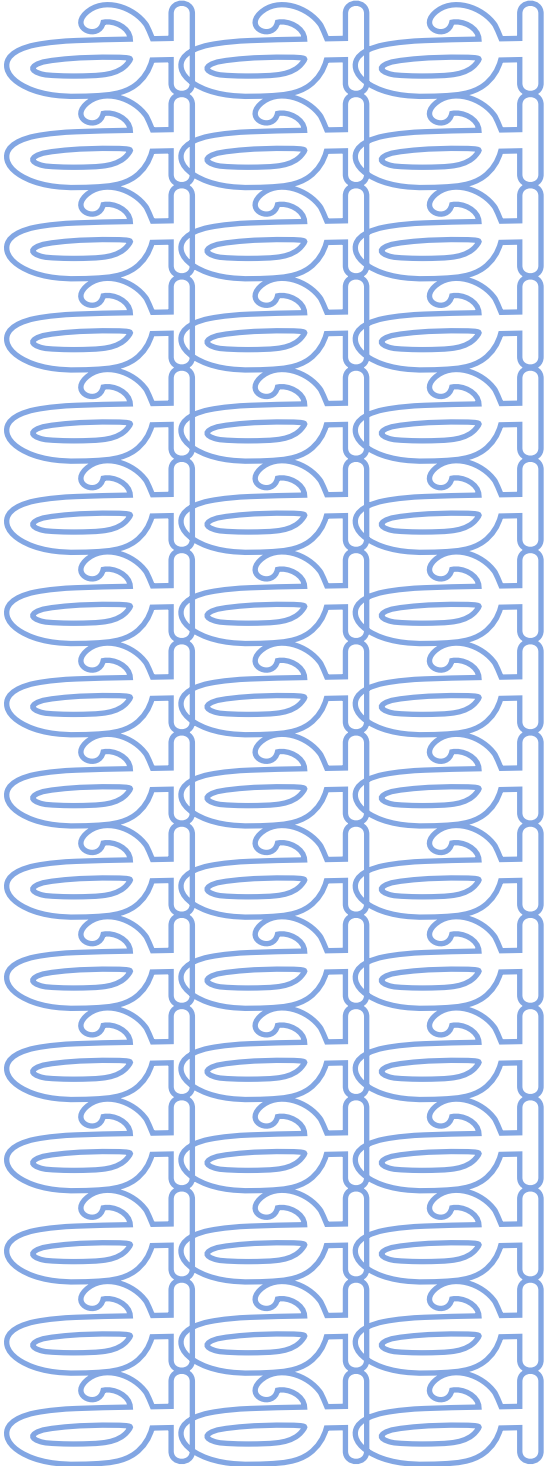
**3. A majority of survey respondents are concerned about politically-motivated moderation and believe that content removals and restrictions are used to silence their voices online, particularly when they speak about politics. A few users also suspect that they experience “shadowbanning,” or a range of opaque, undisclosed moderation decisions by platforms, particularly when they use certain words or symbols commonly used by or associated with the Tamil community.**

**4. Despite a vibrant Tamil computing community, investment in automated moderation in Tamil still falls significantly short due to a lack of accessible resources, will, and financial constraints for smaller social media companies.**





# Tamil as a Low-Resource Language



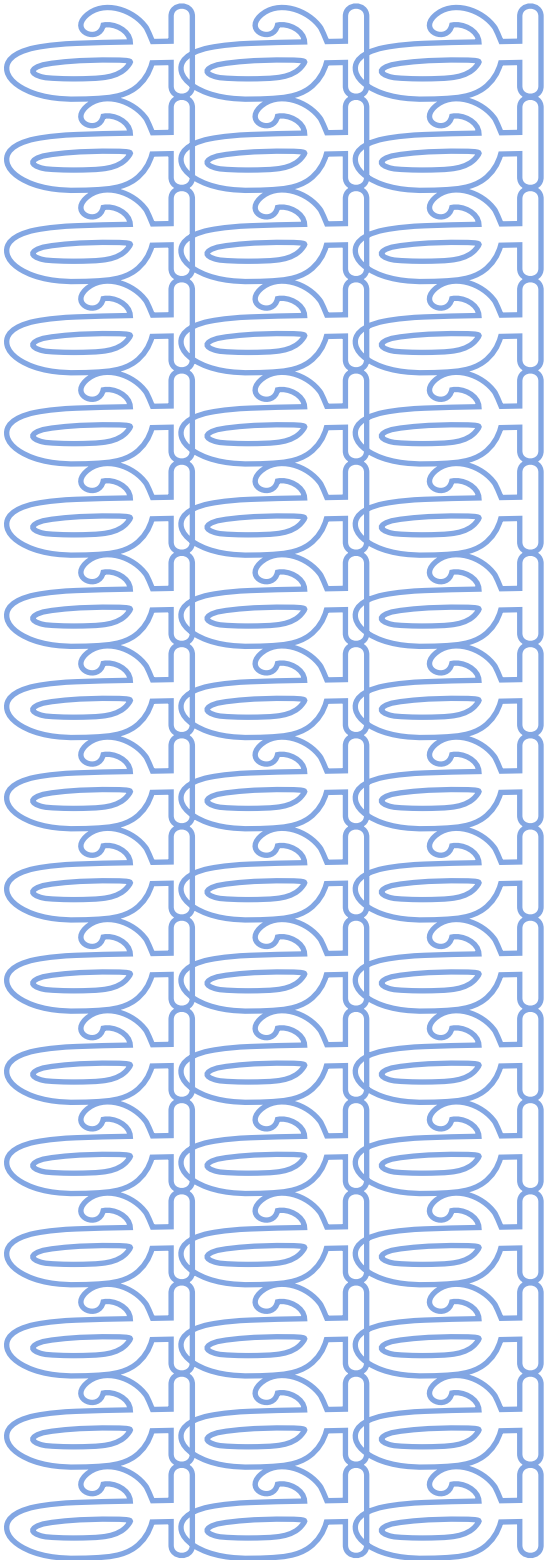
The availability of datasets and digitized text in a particular language, a measure known as “resourcedness” within the computer science field, is critical to train automated tools to enforce content policies and detect content that violates them (Nicholas & Bhatia, 2023). Despite being the 17th most spoken language in the world, Tamil is considered a low-resource language and has significantly fewer high-quality datasets and automated content moderation capabilities than other widely spoken languages. Even amongst other Indic languages, Tamil is a low-resource language; Indo-European languages such as Hindi and Bengali have more digital artifacts than do Tamil or other Dravidian languages (Joshi et al., 2022; Gala et al., 2023).<sup>1</sup> The scarcity of research and datasets in low-resource languages poses challenges in building, testing, and improving automated tools for content moderation, which are predominantly tailored for high-resource languages like English and others commonly spoken in Western contexts (Shahid & Vashistha, 2023; Nicholas & Bhatia, 2023b; Rowe, 2022).

Tamil’s unique and complex form also complicates moderation, particularly automated moderation. Tamil is a diglossic language, meaning that it has multiple variations. Written Tamil tends to have more formalized rules, and often different words, than spoken Tamil. Contemporary spoken Tamil often does not have a standard grammar or lexicon (Nanmalar, Vijayalakshmi & Nagarajan, 2024), and differs depending on the region the speaker is from or even their level of educational attainment (Schiffman, 1996). Sri Lankan Tamil also differs from Tamil spoken in Tamil Nadu and India. Some NLP experts we spoke with argue that Tamil datasets often over-index on Tamil spoken and written by Indian speakers.

Tamil is also an agglutinative language, meaning that the language’s syntax consists of words being added to each other to create another more context-specific word. In this way, Tamil consists of incredibly specific words like கணிதக்கழகப்போராட்டமுயற்சி or “kaṇitakkazakappōrāṭṭamuyar̥ci,” which means “an attempt of struggle at a mathematical institution.” This characteristic complicates digitizing the language, as numerous context-dependent terms are often not represented in datasets.

1 In the papers submitted to the annual Association for Computational Linguistics conference, Hindi was mentioned 16 times and Tamil only once (ACL Rolling Review Dashboard, n.d).

## Tamil as a Historically Politicized Language



Tamil archivists say the language has been historically marginalized and politically contested, particularly in Sri Lanka ([Dassanayake, 2024](#)). In 1956, the passage of the Sinhala Only Act in Sri Lanka made Sinhalese the sole nationally recognized language in the country, impeding the production of Tamil text and even criminalizing the import of Tamil literature ([Anandakugan, 2020](#)). The UN High Commission for Refugees states that language policies put in place dampened the use of Tamil during the civil war which took place between 1983 and 2009 ([UNHCR IRIN, 2012](#)). In 1981, the burning of the Jaffna Library, widely considered one of Asia's largest libraries, destroyed a significant collection of Sri Lankan Tamil writing and artifacts.<sup>2</sup> The former chief librarian of the Jaffna Library said that “six rooms full of material — 97,000 volumes — were turned to ash” ([McCarthy, 2015](#)). As a result, Sri Lankan Tamil has fewer resources than other Tamil dialects.

Tamil speakers in India too assert threats to the proliferation of the Tamil language, and have long countered proposed laws that some say require public institutions to operate in the two dominant national languages, Hindi and English. Tamil speakers argue that these bills erase their linguistic and cultural heritage ([Venkatachalapathy, 2022](#); [Times of India, 2024](#)); language activists have perceived these bills as flattening the nation's polyglot nature, and many have hosted widespread protests ([Subramaniam, 2024](#)).<sup>3</sup>

Some Indian government officials in particular say the use of Tamil in India is endangered, and have thus been resisting efforts related to language policies since the early 20th century ([Sivapriyan, 2025](#); [Venkatachalapathy, 2022](#); [Times of India, 2024](#)). In response, the Tamil Nadu government has sought to nationalize Tamil literature to increase public digitized access to Tamil ([Digital Tamil Studies, n.d.](#)).

Across contexts, the political history of the Tamil language is important to consider to understand Tamil speakers' widespread perception that their language is endangered or mistreated, the related desire to preserve the language both online and offline, and the dearth of high-quality Tamil resources.

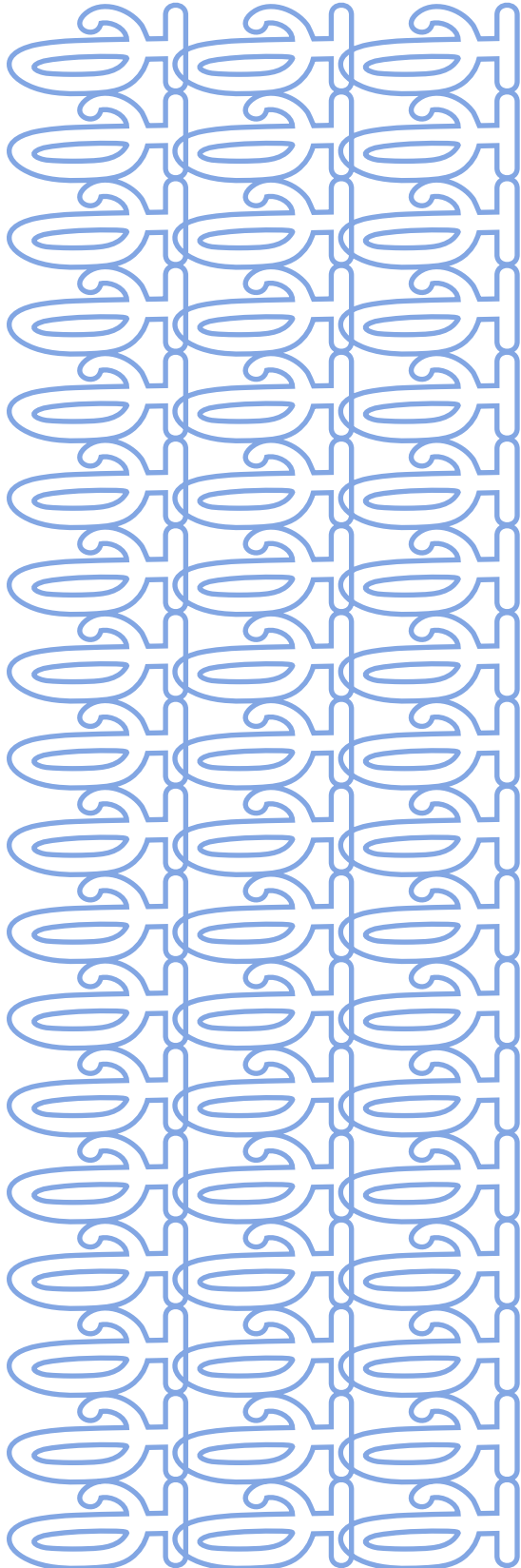
- 
- 2 One Tamil archivist we interviewed said that the burning of the Jaffna Library was a “calculated systematic approach [to] destroying Tamil's cultural identity, heritage, and will to continue to resist.”
  - 3 Political parties and activists have long been opposing these efforts with the first documented acts of protest opposing Hindi-first language policies happening in 1938 and more recently, in 2022 and just last year in September 2024.



The moderation of Tamil content online has been underexplored, particularly with regards to the language's diverse variations and political history. This report utilizes qualitative and quantitative data to investigate the current moderation systems used for Tamil content, and additionally offers recommendations to establish fair and equitable human and automated moderation processes that shape Tamil users' experiences.



# Main Findings



## 1. Tamil speakers use a mix of social media services and online forums in code-mixed, computer-mediated, and transliterated Tamil, sometimes to circumvent moderation.

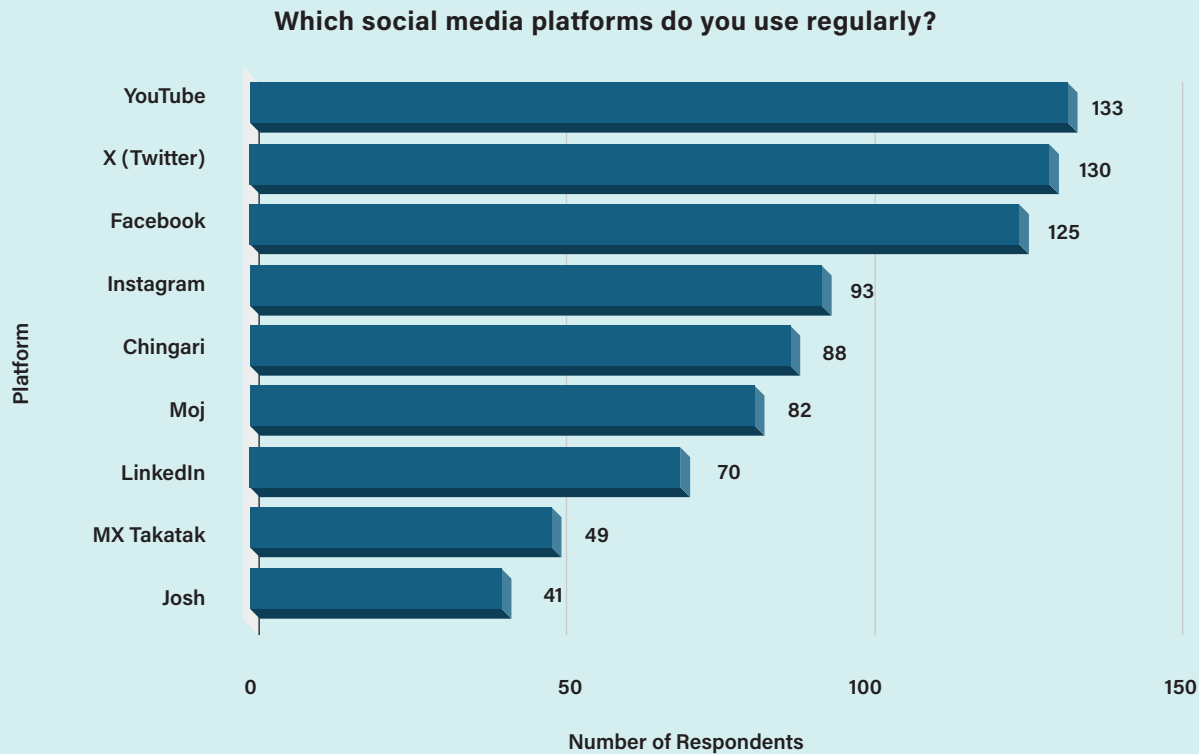
Both U.S.-based and Indian social media platforms are popular among Tamil speakers and Indian Tamil users. Amongst the 147 frequent social media users we surveyed, U.S.-based social media platforms, such as YouTube, X, Facebook, and Instagram, are among the most popular. Survey participants also noted their use of Indian-based platforms in India, with 60% (n=88) of participants reporting using the video-sharing mobile app Chingari. Other applications developed to replace TikTok in India following its ban in June 2020, such as Moj, Josh, and MX Takatak, are used by Tamil speakers (see Figure 1).

Online, Tamil is used in many different ways. Written Tamil is often seen as more formal, while spoken Tamil is considered more colloquial and common. As a result, Tamil posted on social media is often a written form of spoken Tamil with fewer specific rules and norms. Interview participants said Tamil speakers often write in code-mixed terms or messages, including merging Tamil and English to speak in Tamlish or Tanglish<sup>4</sup>. For example, they may say “super-a-irukku” to describe something as great, or commonly switch between Tamil and English in one sentence. Tamlish/Tanglish is particularly common amongst Gen Z and Alpha users, who have evolved meanings for common words. One example is மட்டும் (matter), meaning an incident or conflict, commonly used in a sentence like “konjam matter iruku,” meaning “there’s a bit of a problem.”

The use of Tamlish or Tanglish also includes speech patterns where users use the Latin script to transliterate Tamil speech. Many participants spoke in Tamil online by transliterating the language using Latin characters and writing out the phonetic spellings of Tamil words and sentences. Participants described this practice as a symptom of the ubiquity of internet-connected devices with Latin-script keyboards, and the relative inaccessibility or unavailability of Tamil keyboards.

Consequently, users speaking any variation of Tamil online have developed creative tactics to evade automated moderation systems.

4 Tamlish or Tanglish encompasses users speaking using both Tamil and English words in sentences, and sentences that use the Latin script to transliterate Tamil speech.



▲ **Figure 1.** Number of survey respondents who use social media services in Tamil regularly (daily or weekly) (n=147). Participants were allowed to select multiple options). Source: CDT's online survey (November-December 2024).

In our interviews, participants indicated that they may use code-mixing, transliteration, or “algospeak” to avoid moderation. Algospeak or computer-mediated communication refers to the development and use of code words to evade content removal or down-ranking by moderation systems (Lorenz, 2022; Treem et. al, 2020). In our interviews, an LGBTQ+ activist who routinely faced harassment online said that people used asterisks which they “type[d] for escaping” moderation.

Another participant expressed:

*“They will use the classical word for dog, instead of the colloquial one. Or they will flip the letters of [the] classical word of “dog” so readers will understand but the algorithm will not. Or they will put a “na” or [asterisk] or symbols to circumvent the algorithm.”*  
(Tamil language advocate, November 2024, India)

Another form of circumvention, particularly employed by Tamil users and archivists seeking to share archival Tamil documents from Sri Lanka, involves blurring symbols or cropping posts strategically to evade multimedia moderation techniques. Multimedia moderation techniques use automated tools to analyze and moderate content like images, audio, and video (Thakur & Llansó, 2021). In the aftermath of the Sri Lankan civil war, Tamil users in Sri Lanka and diaspora communities sought to preserve Tamil history and documents online

given the loss of vast volumes of Tamil artifacts in the burning of the Jaffna Library. Many interview participants said they faced content takedowns for sharing archival images or artifacts online from the civil war, particularly when these materials referenced the Liberation Tigers of Tamil Eelam (LTTE).<sup>5</sup> The LTTE, defunct since the end of the war in 2009, has been designated a terrorist organization by some platforms and governments ([Cronin-Furman & Arulthas, 2021](#)). Even when news sites or scholars post content related to the LTTE, they face burdensome moderation. Multiple interview participants pointed to the takedown of the news media site Tamil Guardian's Facebook and Instagram accounts in the aftermath of reporting about the LTTE in Sri Lanka, despite its status as a news organization, as emblematic of this overbroad takedown regime ([Biddle, 2022](#); [Amarsingham & Nandakumar, 2021](#)). Interview participants said that, even when they posted newspaper clippings or documents referencing the LTTE, their posts would be taken down:

*"I used to post these old newspaper clippings from the 1980s and they would censor that. These were clippings about the LTTE or posters or things like that. Things that said 'we will, you know, liberate ourselves, like break chains, that sort of stuff. Not even things like symbols. That would get censored. In response, people would blur posts on Instagram. The poster was in Tamil and they somehow were able to recognize it. I wanted to share historic stuff that [I] was coming across in my studies, or as I was doing research. When I wasn't able to do that, then I was just like 'I don't need to be on this platform for any reason. I didn't really care about it that much. Do you know about the Jaffna library? It was burned down in 1981...hugely traumatizing because a lot of historic stuff was there and people still remember it. In response to that, Tamils created online archives...But you don't see it as much on Instagram and Facebook.'" (Tamil digital rights academic and advocate, December 2024, United Kingdom)*

---

5 The LTTE was designated a terrorist organization by 33 countries in the 1990s after two high-profile political assassinations and several other attempted assassinations. In the years after 2009, with the dissolution of the LTTE after the end of the civil war, a number of countries relaxed their enforcement of this designation. In the last ten years, a number of sitting politicians in Canada and the United Kingdom, where a majority of the Tamil diaspora reside, have attended and taken photos at events with LTTE symbols and flags ([Tamil Guardian, 2023](#)).

## 2. Globalized vs. localized approaches: Social media services and online forums pursue a mix of approaches to moderating content in Tamil.

Through interviews with participants working at either Western or Indian online services, we found that companies pursue a mix of approaches when it comes to Tamil content moderation. Most commonly, companies pursue a “global” or language-agnostic approach to content moderation, as found in an earlier case study in this series (Elsawah, 2024a). However, we also found that some companies adopted a “localized” or language-specific approach when they began offering their service to Tamil users.

### GLOBAL APPROACH

Most U.S.-based large social media companies we spoke with have opted for a global content moderation approach, employing uniform content policies for all users regardless of the language they used or where they were located. This approach consists of recruiting moderators or policy experts without language-specific expertise. One former U.S.-based social media company policy lead referred to their company’s content moderation approach to languages as a “coverage model,” where the question of whether adequate resources were available to moderate content in a given language was mostly considered in times of crisis. They noted:

*“Language would come up when there was a crisis. That would mean either that there was actually a safety risk...there was content on the platform that was being linked to on-ground violence, some kind of conflict breaking on ground and there was dialogue either in civil society or in media. That was when it would come up in terms of reviewing whether or not there was adequate language coverage, whether or not there was classifiers on hate speech.” (Policy lead at a U.S.-based social media company, December 2024, India)*

In the global approach, moderators are hired mostly to do their jobs in English. A participant expressed:

*“Content reviewers are expected to have a certain understanding of the English language up to a certain level. They should be able to understand the policies and make sure that [the policies] have the intended effect.” (U.S.-based social media policy representative, December 2024)*

Sometimes content is automatically machine-translated into English and reviewers are not told what the original language was. Other times, posts are shared with reviewers and Trust & Safety teams, who are asked to rely on machine translation tools to translate pieces of content if they do not speak the language it is presented in, despite machine translation tools being known to be error-prone, particularly in low-resource languages (Nicholas & Bhatia, 2023). A policy lead explained:

*“Content is automatically translated... The platform itself has inbuilt translators and those are known actually not to work well in regional languages at all. If there’s moderation happening in India, Tamil maybe will be translated to English and then the moderator will look at it...the platform’s translation will give it to you in English to the best that it can approximate.” (U.S.-based social media policy lead, December 2024, India)*

When U.S.-based social media platforms do hire native Tamil-speaking content moderators, they do not always task them with reviewing Tamil content. Often, content that is human-reviewed at these large-scale social media companies is given to a moderator at random rather than due to their language proficiency. A policy team member of a U.S.-based social media company said that this practice is intended to ensure consistency in content moderation rules all over the world:

*“If there are regional specifications, that would be included in all content reviewer guidance and available to everyone. The reason for this is that a piece of content, even if it’s coming out of South Asia, it could potentially go to a reviewer in EMEA [which stands for Europe, Middle East, and Africa].” (A policy lead at a U.S.-based social media company, December 2024, India)*

## LOCALIZED APPROACH

A few Indian social media companies and U.S. online forums have adopted localized approaches to moderation of Tamil speech. None of the U.S. social media companies we spoke with employed a localized approach to content moderation. Localized approaches were examined in a previous case study on content moderation in Maghrebi Arabic (Elsawah, 2024a). In this Tamil case, platforms not only alter policies to better suit Tamil contexts, but also give moderators extra agency to communicate with users through blog posts, provide guidance, and share feedback with the company policy teams. Western tech companies which employed localized approaches in the Tamil contexts were more often to be online forums rather than large social media companies.



They employed the local approach by hiring Tamil experts, equipped them with channels such as frequent meetings or messaging spaces to speak with policy teams to create more Tamil language-specific guidance, and developed resources to communicate terms of service to Tamil speakers before offering the service to Tamil speakers. A Tamil moderator talked about how they frequently communicate with policy teams:

*“When the [online forum]’s Indian languages version[s] were launched, all the policies that were available on [the online forum]’s English version were translated to fit into the languages... There were gray areas because most of our policies were translated. That’s the time when you go back and have a discussion to understand if this is something you need to look into and then change the policy.” (Tamil community moderator for an online forum, India, December 2024)*

Some localized content moderation approaches rely heavily on users to flag violations, and use these user reports as trusted signals to inform moderation. To create high-quality user reports, building knowledge and trust within the community is essential. One way community moderators do this is by sharing more specific moderation guidance when a post violates the terms of service. A moderator indicated:

*“Whenever I see that there is a high number of specific policy violations, I’ll make sure to make an announcement [on the service in the form of a blog post] about those violations in policy. Not pointing out people’s names but pointing to generic [posts]. I’ll also send that as a direct message to the user and make them aware... We let all the comments; we trusted the community to report the comments to us if something is not civil, something is triggering or hateful, and we get such reports from the community and we act upon that.” (Online forum moderator, November 2024, India)*

Indian social media companies also localize their content policies, but not always. Largely, Indian technology companies pursue a pan-Indian approach to content moderation, treating speech from all users and regions the same. However, at times, they do localize their policies to account for cultural differences, largely on an ad hoc basis:

*“Tamil is one of the biggest languages on the platform, bigger than Hindi... One of the many reasons for this is the very big entertainment industry...A junior in the team, a Tamil person,*

*wrote those documents and modified those documents to give them a local context...For instance, for the Tamil context, I don't think this was ever written down, but you can't say something about Periyar<sup>6</sup> in the Tamil version of the [platform]...There were certain words in each language that we had muted. Content moderation teams would decide which words are okay and which are not. The Rules could not be enforced equitably. In different languages, certain words would sound very different. In English, the word f\*ck we could use freely. The word in Hindi, we may not use it as freely. I don't know what the word is in Tamil, but again if the word sounds okay, we may allow it." (Indian social media company policy lead, October 2024, India)*

One big challenge for Indian companies is the cost of procuring language-specific moderation. Building automated tools to detect violating content and enforce content policies or hiring reviewers in each language, company representatives said, costs a lot of money. When investing in automated tools for moderation, instances in which companies face potential liability for hosting particular content are prioritized.

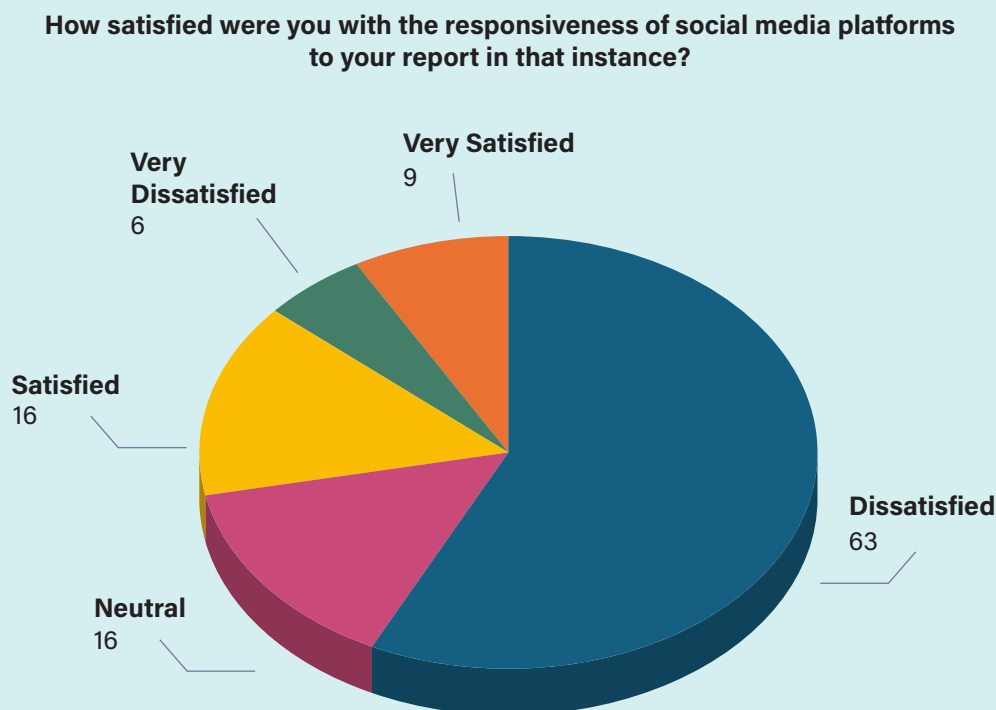
Most survey respondents said that they faced online environments filled with what they perceived as misleading or harmful content. About 81% of survey respondents expressed concern about the proliferation of misleading content, and 82% expressed concern about hate speech. Relatedly, almost 75% of our survey participants indicated that they have at least once reported content online that they perceived as harmful. About 63% of those who had reported violating content indicated that their reports were not helpful and the content was not removed. About 57% were dissatisfied, and 5% were very dissatisfied, with the platform's responses (See Figure 2).

One digital rights advocate expressed their frustration with their content being flagged as harmful while hateful content they reported remained untouched, saying:

*"Whenever I flag to X, I flag to Meta, Instagram, nothing happens. I am doing it because I follow protocol, I follow the*

---

6 Periyar, born Erode Venkatappa Ramaswamy, was an Indian social activist and philosopher known widely as a leading figure of the social justice and anti-caste oppression movement.



▲ **Figure 2.** Survey respondents' perceptions about the effectiveness of reporting systems and processes (number of survey participants who stated that they have reported content to a platform at least once. n=110). Source: CDT's online survey (November-December 2024).

*content moderation policies because I get a lot of flags myself even when I write the smallest thing. I'll get a notice saying 'your post goes against community guidelines.' I don't post offensive stuff, I'm a trainer, a human rights activist, I post very normal stuff, but even for that, I get flagged as a content violation. But when people say 'I'm going to kill you' in Tamil and I flag it, they don't take it down. They say 'We have not found any violation.'...I'm like what? I mean try to translate it...we have given you translations, we've given you curse words, we've given you caste-based slurs in Tamil. We people all over the world have contributed to an annex [or list] and they pass it on to platforms, and they're still not doing anything. When people give threats to me, nothing happens. I post a screenshot, that gets taken down as a 'Content violation.' They just called me a prostitute. They called me those words and I'm just posting a screenshot. But the same [original post], it doesn't [get taken down]. It frustrates you a lot.' (Tamil journalist and digital rights advocate, November 2024, India).*

Interview participants say sexist, homophobic, and caste-based harassment and slurs are rampant on Tamil-speaking online forums. However, interview participants believe that the response by companies differs based on language and who is affected. Referring to the use of bilingual slurs, trolling, and memes, an interview participant said:

*“Misogynist slurs might get flagged and might get taken down, but caste-based slurs never. It never gets taken down. Especially in Tamil. Even in Hindi, if you call someone ‘chapri’ and it gets reported, it will get taken down. But when people use [similar] words in Tamil, especially slurs like ‘para’, which is a slur against women of a particular caste calling them a prostitute, it’s been used against me a lot... these words are never taken down.”*  
(Tamil journalist and digital rights advocate, November 2024, India)

In the specific case of caste-based harassment, interview participants widely believed that social media platforms had neglected to invest in improving caste-based moderation. Research has also found that content moderation policies and processes are often inconsistently enforced when it comes to targeted attacks based on caste ([Kain et. al, 2021](#); [Yatharth, 2025](#); [Soundararajan et al., 2019](#)). In 2019, a study found that 13 percent of hateful content found on Facebook in India consisted of caste-based hate speech including slurs ([Soundararajan et al., 2019](#)).<sup>8</sup>

Groups have repeatedly offered crowdsourced and expert-created slur lists and lexicons to help companies train their systems to detect caste-based attacks.<sup>9</sup> Meta, as one example, discloses on their website that they rely on partnerships to develop slur lists in non-English languages, but how and with whom they engage is less clear. Meta calls these “market-specific lists” created by their regional teams and “ongoing qualitative and quantitative analysis on the language and culture of their region or community” ([Meta, 2022](#)).<sup>10</sup> How and whether these teams engage with

7 A slur whose meaning has evolved with widespread use online and has received more research attention than other slurs ([Mukherjee & Desai, 2024](#)).

8 The author of this study was invited to speak at a technology company about caste in 2022, yet the talk was canceled due to the “polarizing and offensive” nature of the topic. The company’s decision to cancel the talk received significant pushback and media coverage ([Tiku, 2022](#)).

9 One list is created by Tattle, which as it stands is the largest crowd-sourced list of slurs in Indian languages including Tamil with over 600 words ([Tattle, 2024](#)). The team consistently updates the list, including by adding new and different code-mixed terms and misspellings, and makes it available on GitHub under the Open Data License for use by researchers and Trust and Safety teams.

10 As of writing this case study, Meta made changes to its Hateful Conduct policy and seeks to increase reliance on user reports, rather than on proactive detection and moderation, to enforce this new Hateful Conduct policy and other policies.

subject matter experts — and consult open-source lists that are expressly created to aid content moderation, and made available to companies at multiple Trust & Safety arenas such as conferences — is unclear.

The majority of survey respondents said they had reported content before, but some interview participants said that they did not know how to do so. An LGBTQ+ content creator in India said “I don’t know how to seek help from social media.” This finding aligns with prior research that shows that many users, particularly women, gender-diverse, and marginalized users around the world, find it challenging to report content on popular online platforms and find these processes either lacking, inconsistent, or flawed in some way (Sehgal & Nambiar, 2024; Vilk 2023).

In addition, many interview participants attribute targeted hateful speech, and lack of response to the posts they report, to the fact that the posts in question are not in English or that they as a user do not speak English as their first language. Many participants believe that Tamil and caste-based slurs are rarely moderated, and as such, bad actors use them with impunity or with a license to troll. This kind of intersectionality (or the way a person’s multiple identities such as language, caste, class, and gender can lead to unique forms of oppression) is important in understanding the impacts of moderation:

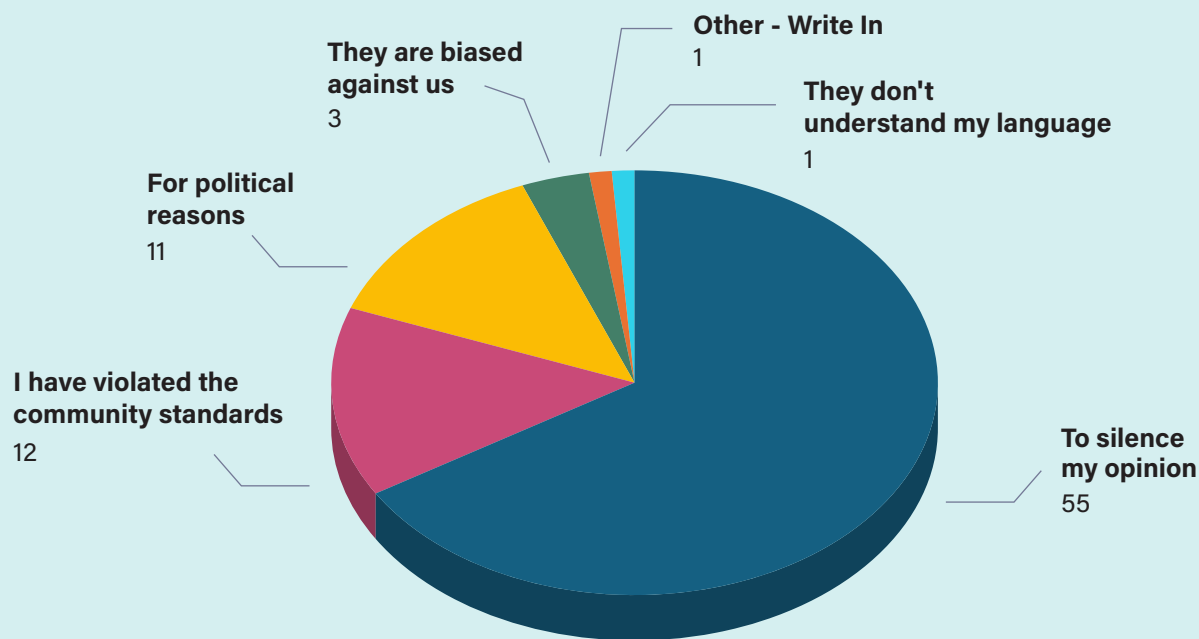
*“You know the treatment of any Indian woman is very different... The treatment of an English-speaking Dalit woman is going to be very different from a Tamil-speaking Dalit woman.”*  
(Tamil journalist and digital rights advocate, December 2024, India)

*“These platforms were not made for us. It was designed for people of a certain class background from my part of the world. It is about class privilege as well. About who gets to go to a school where you are completely taught in English. That’s a privilege.”*  
(Digital Security Trainer, November 2024, India)

### **3. Many users perceive moderation as an effort to “silence” their voices, sometimes on political grounds.**

About 57% of survey respondents indicated that their content had been removed at least once. The majority identified the content as being political in nature. Survey respondents believed that platform

### Why do you think the social media platform removed your content?



▲ **Figure 3.** Reasons participants identified as to why social media platforms removed or restricted their content (number of survey participants who stated that they have experienced content removals or restrictions, n= 83). Source CDT's online survey (November-December 2024).

takedowns and actions perceived to suppress the reach of their posts were taken to “silence” their voices online. These suspicions were pervasive and were repeated in one-on-one interviews with Tamil creators, digital rights advocates, and content moderators.

One longstanding Tamil computing expert and former moderator noted the following when asked about his own experience facing moderation:

*“It doesn’t matter if it’s in Tamil or English, when you are talking against the government the same thing happens. It’s not the language, it’s the bias. The post will be reported as spam and taken down as violating the community guidelines [...] Community guidelines are so vague. It’s not like someone is sitting there and responding to us to understand what exactly went wrong. They’ll just say ‘You violated the community guidelines’ so we read all the guidelines. We don’t understand, what have we violated? I think it’s for their own legal safety.”*  
(Former moderator, October 2024, India)

Another interview participant suspected politically-driven moderation because they posted about Tamils fleeing persecution from Sri Lanka during the civil war. Multiple interview participants pointed to research and human rights advocacy that has long argued that over-moderation in the Sri Lankan context stifles social movements and “obstructs



accountability” (Amarasingam & Nandakumar, 2021; Ethirajan, 2021). One interview participant said that even when people posted images of shopfronts containing the very common name of “Prabhakaran,” which is also the name of the slain leader of the LTTE, their posts would be taken down. One interview participant said that symbols that seemed to be associated with the Tamil nationalist movement were detected and taken down on social media, such as the gloriosa lily.<sup>11</sup> They were not notified whether their post was taken down due to a state request for takedown or proactive moderation, although Sri Lankan and Western media reporting suggest frequent requests for takedowns coming from the government (Constine, 2018; Tamil Guardian, 2024; Mallwarachi, 2024):

*“There was a point during COVID, when there were no mass events. There were a few online activations for November 27th and people were sharing their respect online. This was a period where people weren’t posting the [LTTE] flag or actual pictures of members or uniforms, but were still having content being taken down. The gloriosa lily, I had a friend post a picture with her friend holding lilies saying ‘We remember and resist’ and the post was taken down. It was shocking. How can a flower be violent? How did they find this?” (Sri Lankan Tamil journalist and digital rights advocate, January 2025, United Kingdom)*

A few interview participants also suspected that they experienced opaque moderation, more commonly known as “shadowbanning,” when they used certain terms in Tamil or another language. Shadowbanning is a colloquial term for a practice that encompasses a broad range of undisclosed content moderation actions, such as hiding a user’s posts from other users, removing a user’s handle or posts from search, or ranking users’ content so low in a recommendation system that it is less discoverable (Nicholas, 2022). Prior research showed that, without clear disclosure from the platform about the circumstances under which it might moderate a user’s content without informing them, users develop their own rationale for why their posts face a reduced reach and find ways to circumvent this type of intervention (Savolainen, 2022; Nicholas, 2022). A Tamil political commentator we interviewed emphasized that some of their political content gets less engagement:

---

11 The gloriosa lily is seen as a symbol and the national flower of the Tamil Nationalist movement, which has become synonymous with the Liberation Tigers of Tamil Eelam (LTTE) because it contains all the colours contained in the Tamil Eelam national flag.

*“Especially when I write a lot about the RSS<sup>12</sup>, it goes to the lower feed. I have 54,000 followers but my reach is very limited. Certain posts where you use the word Gaza or Palestine or RSS or BJP, you know you’ll only get...I can close my eyes and say in 10 minutes I’ll get three likes. That’s for sure. That’s how I know it’s been moved to the lower feed.” (Tamil journalist and digital rights advocate, November 2024, India)*

We found that platforms do rely on government input to shape content moderation and often did not disclose this to users. Based on our interviews with former and current platform representatives, platforms often rely on government guidance to shape content moderation in Tamil in the following ways.

First, Indian and Western social media platforms often comply with state requests for takedown or user information with little notice to users. A moderator described that:

*“[Platform] was pretty clear on if a request to takedown a specific content is going to come from a government of a country that states that it is legally obliged to do so, then perhaps [platform] was obliged to remove those content.” (Moderator on social media platform, December 2024, India)*

Second, Indian and Sri Lankan laws related to intermediaries are often interpreted by Indian and Western social media platforms in overbroad and subjective ways resulting in takedown of Tamil content. Laws like the IT Act in India and the Online Safety Act, No. 9 in Sri Lanka require online services hosting user-generated content in the respective jurisdictions to take down illegal or harmful content, with differing degrees of specificity. These requirements have been interpreted in different ways by online services, and often are burdensome to smaller, local intermediaries (Kumar et. al, 2022). This has resulted in Indian platforms, and even Western ones, directly incorporating them into their community guidelines and guidance to moderators, as a head of a policy team at an Indian social media company indicated:

*“The general principle was what came down from the IT Law... You cannot make comments that would create public disharmony*

---

12 The Rashtriya Swayamsevak Singh or RSS is a Hindu nationalist organisation and the parent organization of the Bharatiya Janata Party (BJP), which is a political party, the party which is currently in power federally in India (Britannica, n.d.).

*or discord which means you can't make comments about gods... For instance, Indian law does not allow you to denigrate the national flag, so you can't post content that would denigrate the national flag. Our community standards were largely a subset of that... If something happens on a platform and the police would reach out. If you posted it, you started a riot, the police will ask us for your details. And we'll provide your details... Similarly, there have been various other instances where they have asked for companies to take down things or keep an eye out for things... we'll be a little more proactive about it." (Head of policy, Indian social media company, October 2024, India)*

These laws result in vast takedowns of content, resulting in distrust in platform moderation that causes users to circumvent appeals processes and even content moderation processes altogether. A moderator of an online forum said that disclosure to users was essential to maintaining trust and navigating a tricky balance between compliance and user trust:

*"I may be wrong, but if I'm removing content based on a government's request, and I keep on doing that, the users won't feel trusted that this is a safe space to come and talk. So if I am removing content, I would say to them, 'Because the government said I need to remove it.' If it's the second time, I should at least have a substantive policy to say, 'Hey, look this is some sort of government rule. These are the policies that you need to make sure you don't violate.' You also don't want the government to come back to you again and again and say hey remove this, and this." (Moderator on a small online forum, November 2024, India)*

One employee at a U.S.-based social media company said that distrust in platform moderation results in part from some social media users' perception that companies were aligning themselves with government actors, something that has long been documented by academics, civil society, and media ([Sombatpoonsiri & Mahapatra, 2024](#); [Horowitz & Purnell, 2020](#); [Mirza, 2023](#)).

Finally, Western social media platforms often increased their investments into content moderation capacity only after pressure from government actors. In interviews with Trust & Safety staff, there was a sense that teams do not listen to anyone except for the government. For instance, in the aftermath of the Easter Sunday attacks in 2019, where the Sri Lankan government restricted access to social media platforms

under a state of emergency measure, one former policy employee noted that their company did pull out “breakglass measures” to moderate Tamil content:

*“There are breakglass measures, for example, in the aftermath of the Easter Sunday attacks, the prominent narrative was that the company was blindsided because it didn’t have classifiers in Tamil or Sinhala and there was all this hate speech going unchecked. Which was definitely the case. But at some point, internally, the company did deploy something called ‘Green Lantern,’ which I think is a hate speech classifier, and which was able to contain hate speech in those local languages. There was a lot of secrecy around it because I think companies don’t want it to be known externally that they have these capabilities. So there’s a little bit of ambiguity on whether or not there is an actual resource gap, or whether it was the company’s unwillingness or it doesn’t suit their strategic needs to deploy resources where they might have them.” (Policy Lead at a U.S.-based social media company, December 2024, India)*

In 2018, a year before those breakglass measures were used, companies were accused of failing to moderate instances of incitement of violence amongst Sri Lankan users which led to widespread riots in Sri Lanka. In response to those riots, the Sri Lankan government blocked access to social media services under an emergency order. In 2020, after an investigation commissioned by Meta, one of the companies whose services that order blocked, the company stated that it “recogni[z]ed and apologi[z]ed for the very real human rights impacts that resulted” (Brustein, 2020). The company committed to hiring more staff who spoke Sinhalese and Tamil, and to using detection technology in Sinhalese to protect vulnerable groups including Muslims and Tamils (Facebook Sri Lanka Human Rights Impact Assessment, 2020).<sup>13</sup>

Ultimately, interview participants argued that the real danger is in the growing self-censorship that threatens freedom of speech and the political environment in the region.

*“You have to think a lot before you write and speak. People will automatically self-censor. If the platforms don’t want to censor them, they’ll censor themselves. If one platform is not good, you*

13 Meta also conducted a human rights impact assessment for India but never published it, despite calls for disclosure from international civil society (Access Now, 2022).

*can find another platform. Or you can have your own website or your own server. The platforms taking you down is actually a matter of inconvenience. The real danger is the democratic climate of the country. We are putting ourselves [at] risk. More than the platforms themselves, the real risk is the climate, the democratic climate of the country that you're operating in."*  
(Former moderator, October 2024, India)

#### **4. Despite advancements in Tamil NLP research and content analysis capabilities for Tamil, social media companies are slow in adopting automated moderation tools for Tamil — due in part to low financial investment and lack of engagement with Tamil NLP experts.**

The Tamil community has a rich and tight-knit computing tradition, and many who are involved in community moderation or open-source knowledge-sharing have been building and discussing moderation since at least the early 2000s. Several Tamil computing experts we spoke with began as moderators or community administrators, and are familiar with the delicate balance of encouraging engagement while fostering healthy online communities:

*"At the time, there was a really rich ecosystem of Tamil blogs. Compared to any other Indian language, there was an early Tamil blog community. There were blog aggregators developed by Tamil developers that were popular. I'm talking about 2003, 2002, 2004. This blog community was all on the open web and thriving...I would say Tamil computing-related community is one of the pioneering communities in the context of India. In terms of organized mass contributors working on something."*  
(Tamil NLP expert and former moderator, November 2024, India)

Tamil moderators and computing experts are also driven by immense passion for their language. One computing expert spoke about his identity and experience fleeing persecution in the Civil War as informing his work, and pointing to the importance of investing in Tamil proficiency:

*“A lot of the work that we do in the sense that we need to preserve the language resources, the cultural heritage, outside of traditional sorts of institutions like archives. So, the Noolaham<sup>14</sup> started as a sort of diasporic effort to digitize. Noolaham is a project to digitize printed materials and histories of multimedia documentation. Digital technologies enabled us to make resources available widely.” (Tamil Moderator and NLP expert, October 2024, Canada)*

Despite the robust and vibrant community of Tamil moderators and computing enthusiasts, Western social media companies’ engagement with Tamil-language computing experts has been scant, and the language has remained low-resource. This is particularly true in terms of the development of automated content moderation systems.

Automated content moderation systems are generally important tools in the content moderation toolkit, given the scale at which online services host content (Duarte & Llansò, 2017). Yet, online services do not invest in building automated systems in Tamil due to the sparse availability of training data, the cost of procuring these tools (as one Indian social media company noted), or the low willingness to invest in Indic languages. Experts told us during a Tamil NLP roundtable that existing content moderation and sentiment analysis systems often fall short due to the limited availability of datasets that accurately reflect the linguistic complexity of the language, the increased frequency with which Tamil speakers use Tanglish or Tamlish and algo-speak online, and the unique contextual-nature of the language. Additionally, few automated content moderation systems are tested against natively-created evaluation metrics.

One Tamil expert says that sparse company engagement with Tamil experts is a recent trend, and that he was previously engaged by one U.S.-based social media company to translate the services’ interface.<sup>15</sup> In contrast, a Tamil fact-checker we talked to noted that, when they asked a company they fact-checked for to improve its provided automated tools, the company responded as such:

14 Archivists say that digital open-access archives may hold the key to more sustainable methods of preserving Tamil artifacts. Noolaham Foundation is one such archive, serving as a free internet library that provides access to information sources and “fosters knowledge-based development” in Sri Lanka. It contains over 100,000 items, including books, magazines, newspapers, pamphlets, manuscripts, and more (Noolaham Foundation, n.d; Daniel, 2016).

15 Translated versions of this service were previously crowdsourced by experts, a program that is suspected to have been deprecated according to researchers (u/diegomrosa, 2023).



*“For example, four years back today, [Social media company] tried to do translation accuracy language tools and then transcription. Transcription tools [helped] subtitling in English to Tamil but four years back [there were] no tools for subtitling from English to Tamil. [I learned about] one tool from California. I [sent] a mail to that concerned company. I [asked] why not support in Tamil. That answer is my eye-opener. They said, ‘Tamil language [among the] overall world population [is] under 2%. Why [should] our company manufacture one tool for under 2% of the people? If you want, you can make it for that.’”*  
*(Tamil fact-checker, December 2024, India)*

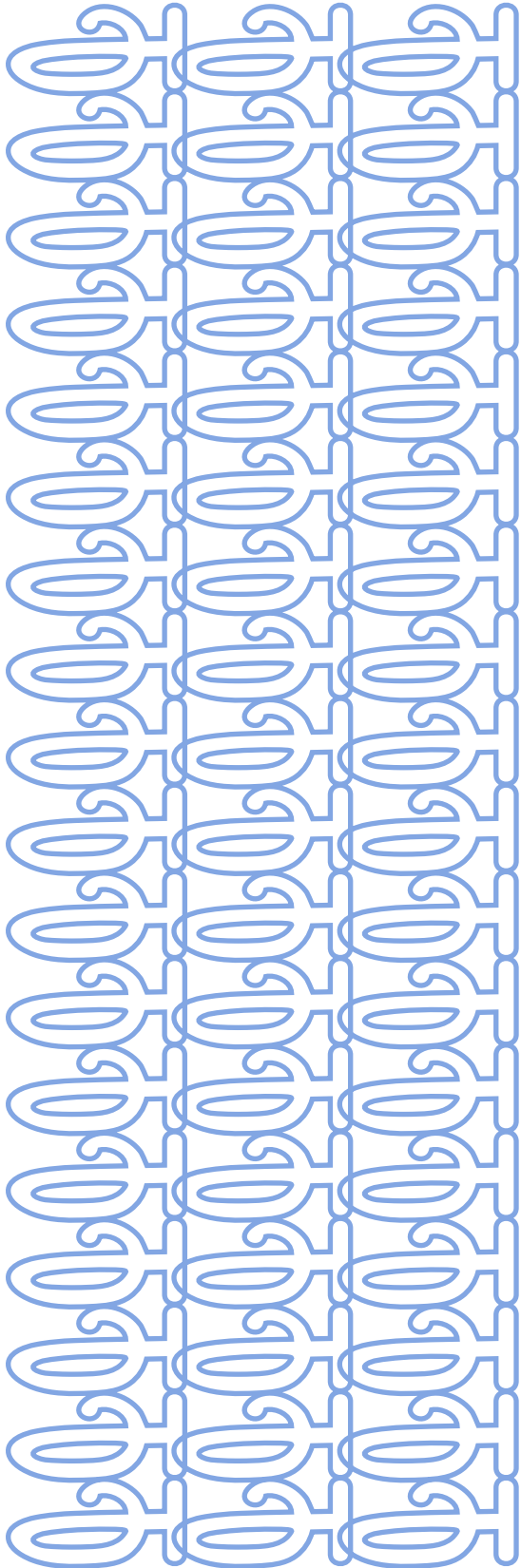
Several Tamil speakers and NLP experts have invested personally or sought grants from state governments to improve the availability of data in Tamil and build automated tools in the language. A number of research consortia and projects, such as the Center for Tamil National Language Processing Research, focus expressly on building and improving automated technologies in Tamil accompanied by efforts led by Microsoft Research and Project Vaani. There are also multiple arenas to advance state-of-the-art Tamil language technologies such as the Association for Computing Linguistics’ DravidianLangTech workshop and the International Conference on Tamil Computing. Initiatives like AI for Bharat and Karya, dedicated to improving Tamil representation among Indic languages in training datasets, exist too. Leading academics like Ruba Priyadharshini and Bharati Chakravarthi are repeatedly cited as leading academics who have studied the use of Tamil online, and built corpora of Tamil text to facilitate sentiment analysis and hate speech detection in Tamil ([Chakravarthi et. al, 2020](#)). These are just a few examples of avenues where Tamil NLP experts said social media companies can look to improve their Tamil tooling.

Interview participants identified many priority areas for greater company investment of NLP funds. The first was research to better understand evolving Tamil speech patterns, particularly online. While the Tamil Nadu government has kickstarted its own initiative to nationalize and digitize Tamil documents, aiming to improve the resourcedness of the language online, more investment into understanding online Tamil speech and building research-based automated technologies for moderating Tamil content is crucial to improving Tamil speakers’ experiences with online moderation.

Finally, developing benchmarks and frameworks to test automated systems is essential to ensure automated moderation technologies work across nationality, caste, gender, and class. Promising initiatives like BHASA and Pariksha have already cropped up. Since models should be informed by a multifaceted understanding of the Tamil language, it is crucial to support more efforts to integrate varying annotator perspectives through the use of participatory methods to develop benchmarks ([Arora et. al, 2024](#); [Gordon et al., 2022](#)).



# Recommendations



## 1. Ensure Tamil content moderation guidance accounts for the diversity of Tamil speech across nationality, gender, caste, and class, by increasing engagement with subject matter experts when developing and reviewing policies.

Interview participants repeatedly noted that content moderation often failed to account for the different ways Tamil speakers communicate online, particularly when companies pursued a global approach to content moderation. Many also believed that moderation decisions didn't appropriately incorporate considerations of different socioeconomic status, gender, sexual orientation, and caste. Study participants felt as though they faced poor moderation because they didn't speak English or because platforms had a higher threshold of tolerance for harmful speech in non-Western contexts. As other researchers have identified, platforms should ensure that moderation guidelines are sensitive to the differing ways that harassment and harm manifest to different users, and don't burden the ability of marginalized users to call out harassment they receive ([Diaz, 2023](#)). Companies should reconsider applying a uniform approach to content moderation without considering unique linguistic nuances, given the potential for that approach to burden the ability of Tamil speakers and other users to express themselves in their chosen language. Even when pursuing a uniform approach to content moderation across language and region, platforms should engage language and subject matter experts to understand the different ways policies are interpreted or situated within Tamil-language contexts in India, Sri Lanka, and the diaspora. Doing so will help platforms ensure that violating, harassing speech does not target marginalized users, and that speech by marginalized users isn't erroneously suppressed online.

These efforts are particularly important given that companies are rolling back their already limited Trust & Safety investments ([Goggin, 2024](#)). Companies are opting to use more automated solutions to fill this gap in Trust & Safety capacity, yet as we've demonstrated, Tamil content moderation systems currently have many gaps and limitations. Engaging with Tamil experts including language computing experts and subject matter experts, who best understand how harms manifest and how Tamil is used online, can help companies improve automated

technologies and content moderation processes as a whole. For example, the way gender-based harms show up in Indian, and more specifically Tamil contexts, differs a great deal from other Indic languages or non-Asian contexts ([Aneja et al., 2024](#)).

## 2. Improve enforcement of harassment policies and invest more in capacity for reporting channels in Tamil.

In interviews, reporting tools to flag harassment or hate speech in Tamil were found to be insufficient, lacking, or difficult to use by Tamil speakers. Prior research conducted by groups familiar with the Tamil context also found discrepancies and shortcomings with reporting flows, particularly in the South Asian context and for reporting nuanced cases of harassment within marginalized communities ([Sehgal & Nambiar, 2024](#)). First and foremost, platform policies are not always accurately translated into other languages including Tamil, limiting users from knowing what they can report and how ([Localization Lab & Internews, 2023](#)). Additionally, users are also limited from reporting multiple posts although showing a pattern of targeted replies is sometimes necessary to demonstrate harassment. Users can only report content from an available list of categories, which often do not adequately describe many types of harassment in the Global South ([Sehgal & Nambiar, 2024](#)). Moreover, even when users succeed in reporting, they often get no response from platforms.

Platforms should make sure all user reports are responded to in the chosen language of the user, and in a timely fashion that both reflects the scale on which the platform operates and acknowledges the impact repeated harassment has on users' mental health and well-being ([Thakur & Hankerson, 2022](#)). Particularly in light of current proposals to use large language models (LLMs) to offer better explanations about why a post was or was not taken down, companies should ensure that all steps in reporting flows are tested by native speakers to ensure they are accessible to all users, no matter which language they speak ([Digital Trust & Safety Partnership, 2024](#)).

### **3. Supplement the use of machine translation tools with adequate language proficiency, either by hiring Tamil translators or contracting their services.**

Moderators and moderation teams often rely on machine translation technologies to translate Tamil content into English before it is moderated. Yet, researchers and journalists have found that these automated translation tools are often error-prone, particularly in low-resource languages such as Tamil, and often neglect words, make up entirely new ones, and even completely misunderstand a term when translating Tamil content due to a paucity of Tamil data to train and test these models ([Choudhary et al., 2018](#); [Ramesh et al., 2020](#); [Deck, 2023](#)). Moderation teams looking to use translation tools should consider using natively-created corpora to supplement or improve in-house translation tools, such as [Vaani](#). They should also ensure that content is reviewed by native language speakers, particularly when it comes to borderline content and appeals, so as not to lose the essential context required to parse a post.

### **4. Engage with Tamil NLP researchers and use their research to improve Tamil moderation tools, including by adopting their open source lexicons for hate speech and caste-based speech, and using natively-sourced benchmarks that better understand Tamil speech.**

Automated content moderation and translation technologies are increasingly used to moderate speech, particularly by platforms that offer services at scale. Yet, they fall short in many ways due to the scarcity of high-quality training datasets and resources like benchmarks and evaluation metrics to test if these systems are working as intended ([Nicholas & Bhatia, 2023](#)). Trust & Safety teams are often resource-strapped, and may not be able to build such resources to train and test these systems. However, a number of researchers have built open source lexicons and culturally-specific benchmarks. [Tattle's Uli](#), for instance, is

an open source crowdsourced lexicon of slurs against gender, caste, and other marginalized identities across Tamil, Hindi, and Indian English. Platforms can and should create channels for Trust & Safety teams to engage with the experts who are building these resources, and can help identify, vet, and make use of available datasets to improve Tamil moderation.

Using these natively created lexicons and benchmarks is essential because existing machine-translated datasets and benchmarks are often imperfect proxies to assess the performance of content moderation systems. Researchers say that, while resources like the Massive Multitask Language Understanding (MMLU) benchmark — which contains 16,000 questions across fields — are widely relied upon, they are often not sufficient to gauge the performance of systems used to parse content and context of non-Western speech. For example, while the MMLU benchmark may successfully evaluate a model's performance in answering a question about American law, it may not assess whether a model can answer a question about the capital of Tamil Nadu (in either Tamil or English).

An example of an Indic-specific benchmark that Trust & Safety teams can use is the IndicMMLU-Pro, which establishes a new standard for evaluating models, including content moderation tools and translation tools, in specific localized domains in Tamil and other Indic languages ([KJ et al., 2025](#)). Others include BHASA, a benchmark for Southeast Asian contexts in which Tamil is also spoken widely, and Pariksha, a combination of human and machine-translated benchmarks in several Indic languages produced as part of a Microsoft research project ([Leong et al., 2023](#); [Watts et al., 2024](#)).

Finally, Trust & Safety teams should consider engaging with NLP researchers to use participatory methods when developing or using corpora to test Tamil-language moderation tools. Trust & Safety teams regularly bemoan the challenge of training and testing automated systems, particularly when data annotators or evaluators disagree with one another about what a word means or whether a piece of content is offensive or not. However, researchers say there may be ways to integrate these dissenting voices and disagreements into datasets that can then be used to train models to parse context and complex speech ([Gordon et al., 2022](#); [Arora et al., 2023](#)). This approach can also help ensure that datasets represent the myriad ways users speak and transliterate Tamil.



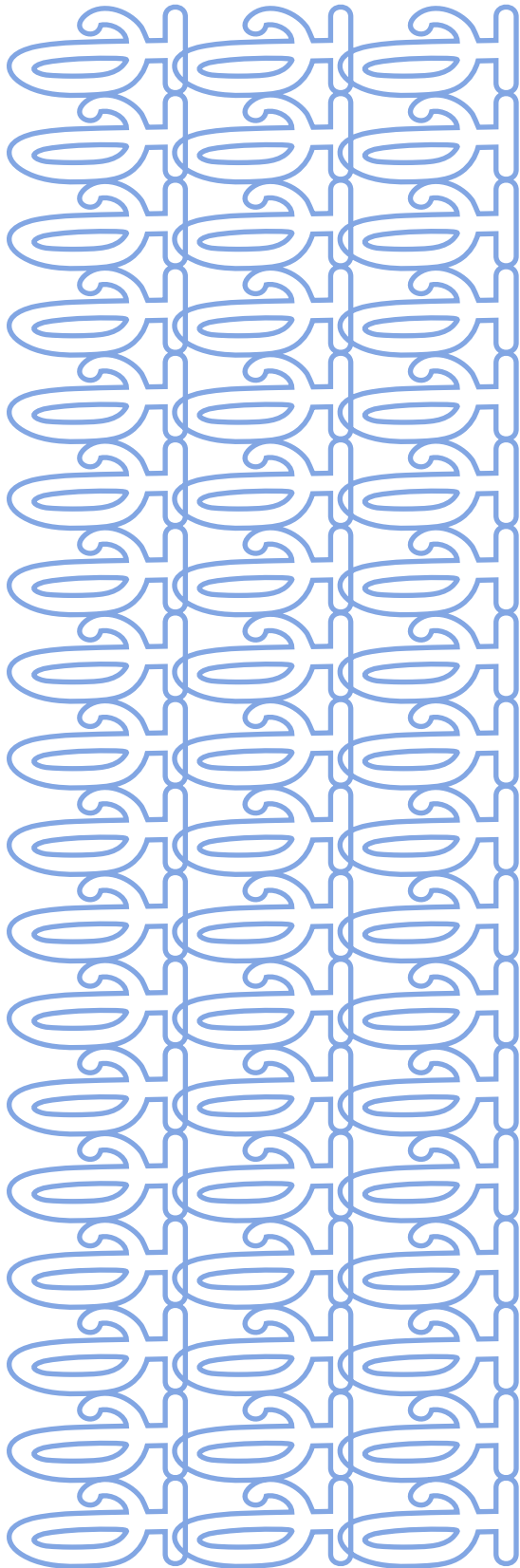
## **5. Disclose to users when content moderation has happened and why, including instances of reduced reach and government requests for takedown, user data, and moderation changes.**

Platforms may moderate content and reduce the reach of posts for any number of reasons, yet users deserve to know when this has happened and why, particularly when decisions are made in coordination with or at the behest of their democratically elected officials. Users can glean a lot from clear disclosure from platforms, including but not limited to: learning how involved their state has been in guiding content moderation; how often or in what way a company interacts with government actors; how to adhere better to platform content moderation rules in the future; making more specific appeals to platforms and oversight boards when they feel like a moderation decision has misunderstood their posts; and knowing when their speech has been curtailed by their government and holding their leaders accountable. Disclosure by companies to users about when and why a post has been moderated is also a critical part of adhering to the [UN Guiding Principles on Business & Human Rights](#) (2011), which technology companies are party to, and more broadly promoting users' human rights online.

Platforms can help users in this endeavor by disclosing how a post was moderated, such as using automated means, through a government request for takedown, or in response to a user request. Platforms can also disclose to users what specific policy their post violated in order for users to learn more about the policy. And finally, disclosure should include a clear channel for users to appeal a platform decision.

In the particular case of government requests, platforms can enable users to better protect their rights by making the nature and text of the government request available. In the past, companies have published these requests in aggregate through transparency reports and public repositories to enable oversight, citizen participation, and research into how governments work with companies. One example of this is the [Lumen database](#), an independent project housed at Harvard University's Berkman Klein Center, where platforms can voluntarily submit a specific government request for takedowns or user data. Making government requests received by platforms public in forums like the Lumen database or the like should be a critical part of the Trust & Safety efforts to engender trust and transparency.

# Appendix



## Methods and Data Collection

To study the state of content moderation in Tamil, we used a mixed-methods approach, combining qualitative and quantitative methods. First, we conducted semi-structured interviews with 17 experts to assess the information environment, learn more about the state and perception of content moderation, and understand firsthand how content moderation works and does not work. We spoke with Trust & Safety and policy professionals at Western and Indian companies, moderators tasked with reviewing speech from South Asia or the Asia Pacific region, digital rights experts and academics, and Tamil-language content creators.

Most interviews were conducted online, except for one that was conducted in person during a field visit in Bangalore, India. The interviews were conducted between October 2024 and January 2025. The interviews were all conducted in English, with some discussion in Hindi, and some examples given in Tamil that the interview participant translated into English for the research team. Field notes were taken during the interviews and recordings of the interviews were later transcribed and analyzed to find common themes.

In addition to the interviews, we conducted an online survey to understand the experience of Tamil-speaking users who used social media frequently. The survey was circulated by our partner, Centre for Internet & Society, based in Bangalore, as well as by interviewees who had previously been part of Tamil online communities. The survey was also circulated in a Tamil subreddit. The survey asked questions about users' trust in social media platforms, their experiences facing moderation, and their perceptions of moderation. The Alchemer platform was used to distribute the online survey from November 7th to December 15th, 2024. A modest honorarium of US\$10 was offered for participating. The survey was available in English and in Tamil. We targeted participants who were frequent users of online services. About 90% of our survey participants indicated that they use social media on a daily basis. The majority of our sample were individuals from India (n=138) with only a few from Sri Lanka. To address this limitation, we conducted interviews with Sri Lankan Tamil users and experts and invited them to participate in our roundtable.

Finally, on January 8, 2025, we held a roundtable with 10 computer science experts familiar with building, evaluating, and annotating datasets for machine learning technologies in Tamil. The roundtable discussion was held online and consisted of experts from Canada, India, the United Kingdom, and the United States, and sought to address the challenges those experts faced in data collection, annotation, evaluation, and solutions they found to address these challenges.

## References

- Access Now. (2022). *Meta must disclose India's Human Rights Impact Assessment*. Access Now. <https://www.accessnow.org/press-release/meta-india-human-rights-impact-assessment/> [perma.cc/7X2M-8YZT]
- ACL Rolling Review Dashboard,. (n.d.). *ARR Dashboard*. Languages Mentioned in Paper Abstracts. Retrieved February 24, 2025, from <https://stats.aclrollingreview.org/submissions/linguistic-diversity/> [perma.cc/U2FM-3UEN]
- Amarasingam, A., & Nandakumar, T. (2021, May 14). Social Media Platforms are Silencing Social Movements. *Tech Policy Press*. <https://techpolicy.press/social-media-platforms-are-silencing-social-movements> [perma.cc/7BQ3-JE4L]
- Anandakugan, N. (2020, August 31). *The Sri Lankan Civil War and Its History, Revisited in 2020*. Harvard International Review. <https://hir.harvard.edu/sri-lankan-civil-war/> [perma.cc/7V6J-9SGB]
- Aneja, U., Gupta, A., Jain, A., & John, S. (2024). *From Code to Consequence: Interrogating Gender Biases in LLMs within the Indian Context*. Digital futures lab. <https://www.notion.so> [perma.cc/6VA2-H73U]
- Arora, A., Jinadoss, M., Arora, C., George, D., Brindaalakshmi, Khan, H. D., Rawat, K., Div, Ritash, Mathur, S., Yadav, S., Shora, S. R., Raut, R., Pawar, S., Paithane, A., Sonia, Vivek, Priscilla, D., Khairunnisha, ... Prabhakar, T. (2024). *The Uli Dataset: An Exercise in Experience Led Annotation of oGBV (arXiv:2311.09086)*. arXiv. <https://doi.org/10.48550/arXiv.2311.09086> [perma.cc/H6CV-UEG6]
- Biddle, S. (2022, January 19). *Facebook's Tamil Censorship Highlights Risks to Everyone*. The Intercept. <https://theintercept.com/2022/01/19/facebook-tamil-censorship-sri-lanka/> [perma.cc/Q2TX-KYLK]
- Brustein. (2020, May 12). Facebook Apologizes for Role in Sri Lankan Violence. *Bloomberg*. <https://www.bloomberg.com/news/articles/2020-05-12/facebook-apologizes-for-role-in-sri-lankan-violence> [perma.cc/3VUR-KCLY]
- Chakravarthi, B. R., Muralidaran, V., Priyadarshini, R., & McCrae, J. P. (2020). *Corpus Creation for Sentiment Analysis in Code-Mixed Tamil-English Text (arXiv:2006.00206)*. arXiv. <https://doi.org/10.48550/arXiv.2006.00206> [perma.cc/7NXF-YW5J]
- Choudhary, H., Pathak, A. K., Saha, R. R., & Kumaraguru, P. (2018). Neural Machine Translation for English-Tamil. In O. Bojar, R. Chatterjee, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, C. Monz, M. Negri, A. Névéol, M. Neves, M. Post, L. Specia, M. Turchi, & K. Verspoor (Eds.), *Proceedings of the Third Conference on Machine Translation: Shared Task Papers* (pp. 770–775). Association for Computational Linguistics. <https://doi.org/10.18653/v1/W18-6459> [perma.cc/N24Q-6TAS]
- Constance, J. (2018). Facebook reveals 25 pages of takedown rules for hate speech and more | TechCrunch. *Tech Crunch*. <https://techcrunch.com/2018/04/24/facebook-content-rules/> [perma.cc/4JDD-UFL9]
- Cronin-Furman, K., & Arulthas, M. (2024). How the Tigers Got Their Stripes: A Case Study of the LTTE's Rise to Power. *Studies in Conflict & Terrorism*, 47(9), 1006–1025. <https://doi.org/10.1080/1057610X.2021.2013753> [perma.cc/3AFF-VFJR]

- Daniel, S. (2016, February 18). *Sri Lankan Tamils around the world have built an online library to replace one torched in 1981*. Scroll.In. <http://scroll.in/article/802923/sri-lankan-tamils-around-the-world-have-built-an-online-library-to-replace-one-torched-in-1981> [perma.cc/DHR9-GV24]
- Deck, A. (2023, September 6). We tested ChatGPT in Bengali, Kurdish, and Tamil. It failed. *Rest of World*. <https://restofworld.org/2023/chatgpt-problems-global-language-testing/> [perma.cc/GFD3-YBJR]
- Diaz, A. (2023). *Online Racialization and the Myth of Colorblind Content Policy* (SSRN Scholarly Paper 4509790). Social Science Research Network. <https://doi.org/10.2139/ssrn.4509790> [perma.cc/5GS2-KNN8]
- Digital Tamil Studies. (n.d.). *Tamil Nationalized and Public Domain Books Collection*. Retrieved April 5, 2025, from <https://tamil.digital.utsc.utoronto.ca/61220/utsc35335> [perma.cc/X59S-YCVC]
- Digital Trust & Safety Partnership. (2024). Best Practices for AI and Automation in Trust & Safety. [https://dtspartnership.org/wp-content/uploads/2024/09/DTSP\\_Best-Practices-for-AI-Automation-in-Trust-Safety.pdf](https://dtspartnership.org/wp-content/uploads/2024/09/DTSP_Best-Practices-for-AI-Automation-in-Trust-Safety.pdf) [perma.cc/3JU2-PW9B]
- Duarte, N., & Llansó, E. (2017). *Mixed Messages? The Limits of Automated Social Media Content Analysis*. <https://cdt.org/insights/mixed-messages-the-limits-of-automated-social-media-content-analysis/> [perma.cc/Z5W3-B6ZQ]
- Elsawah, M. (2024a, September 27). Moderating Maghrebi Arabic Content on Social Media. *Center for Democracy and Technology*. <https://cdt.org/insights/moderating-maghrebi-arabic-content-on-social-media/> [perma.cc/XL8Y-BJ8H]
- Elsawah, M. (2024b, December 12). Moderating Kiswahili Content on Social Media. *Center for Democracy and Technology*. <https://cdt.org/insights/moderating-kiswahili-content-on-social-media/> [perma.cc/J8CM-MMDE]
- Ethirajan, A. (2021, March 23). *UN to collect evidence of alleged Sri Lanka war crimes*. <https://www.bbc.com/news/world-asia-56502221> [perma.cc/U9AY-AE8L]
- Facebook Sri Lanka Human Rights Impact Assessment. (2020). *Facebook Response: Sri Lanka Human Rights Impact Assessment*. Facebook. <https://about.fb.com/wp-content/uploads/2021/03/FB-Response-Sri-Lanka-HRIA.pdf> [perma.cc/XKC5-HWCL]
- Gala, J., Chitale, P. A., AK, R., Gumma, V., Doddapaneni, S., Kumar, A., Nawale, J., Sujatha, A., Puduppully, R., Raghavan, V., Kumar, P., Khapra, M. M., Dabre, R., & Kunchukuttan, A. (2023). *IndicTrans2: Towards High-Quality and Accessible Machine Translation Models for all 22 Scheduled Indian Languages* (arXiv:2305.16307). arXiv. <https://doi.org/10.48550/arXiv.2305.16307> [perma.cc/CG63-JW4Q]
- Goggin, B. (2024, March 29). *Big Tech companies reveal trust and safety cuts in disclosures to Senate Judiciary Committee*. NBC News. <https://www.nbcnews.com/tech/tech-news/big-tech-companies-reveal-trust-safety-cuts-disclosures-senate-judicia-rcna145435> [perma.cc/R92J-P5BD]
- Gordon, M. L., Lam, M. S., Park, J. S., Patel, K., Hancock, J. T., Hashimoto, T., & Bernstein, M. S. (2022). Jury Learning: Integrating Dissenting Voices into Machine Learning Models. *CHI Conference on Human Factors in Computing Systems*, 1–19. <https://doi.org/10.1145/3491102.3502004> [perma.cc/FX9T-FZ53]

- Horwitz, J., & Purnell, N. (2020, August 30). Facebook Executive Supported India's Modi, Disparaged Opposition in Internal Messages. *Wall Street Journal*. <https://www.wsj.com/articles/facebook-executive-supported-indias-modi-disparaged-opposition-in-internal-messages-11598809348> [perma.cc/S4ZK-TCQX]
- Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 6282–6293). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2020.acl-main.560> [perma.cc/9F29-N9PM]
- Kain, D., Narayan, S., Sarkar, T., & Grover, G. (2021). *Online caste-hate speech: Pervasive discrimination and humiliation on social media*. Centre for Internet and Society. <https://cis-india.org/internet-governance/blog/online-caste-hate-speech-pervasive-discrimination-and-humiliation-on-social-media> [perma.cc/V6PT-P7NE]
- KJ, S., Kumar, A., Balaji, L., Kotecha, N., Jain, V., Chadha, A., & Bhaduri, S. (2025). *IndicMMLU-Pro: Benchmarking Indic Large Language Models on Multi-Task Language Understanding* (arXiv:2501.15747). arXiv. <https://doi.org/10.48550/arXiv.2501.15747> [perma.cc/D6J2-XAZK]
- Kumar, R., Thanugonda, K., & Guha, D. (2022, December 18). *A Framework for Intermediary Classification in India—The Quantum Hub*. <https://thequantumhub.com/a-framework-for-intermediary-classification-in-india/> [perma.cc/EHD2-AXF6]
- Leong, W. Q., Ngui, J. G., Susanto, Y., Rengarajan, H., Sarveswaran, K., & Tjhi, W. C. (2023). *BHASA: A Holistic Southeast Asian Linguistic and Cultural Evaluation Suite for Large Language Models* (arXiv:2309.06085). arXiv. <https://doi.org/10.48550/arXiv.2309.06085> [perma.cc/LBC6-X648]
- Localization Lab and Internews. (2022, November 11). “Wait, Who’s Timothy McVeigh?”: Why localizing tech policies has to start with communities. Localization Lab. <https://www.localizationlab.org/blog/2022/11/11/wait-whos-timothy-mcveigh-content-policies-review> [perma.cc/RNN5-Y8NB]
- Lorenz, T. (2022, April 8). Internet ‘algospeak’ is changing our language in real time, from ‘nip nops’ to ‘le dollar bean.’ *Washington Post*. <https://www.washingtonpost.com/technology/2022/04/08/algospeak-tiktok-le-dollar-bean/> [perma.cc/RM43-BHRU]
- Lumen. (n.d.). [Dataset]. Retrieved February 25, 2025, from <https://lumendatabase.org/pages/about> [perma.cc/6LZV-VGPD]
- Mallwarachi, B. (2024). Sri Lanka passes bill allowing government to remove online posts and legally pursue internet users. *Associated Press*. <https://apnews.com/article/sri-lanka-internet-bill-freedom-of-expression-wickremesinghe-527e0195a8f8f76573aa8d7185562b12> [perma.cc/7KEL-D8JL]
- McCarthy, J. (2015, August 19). Up From The Ashes, A Public Library In Sri Lanka Welcomes New Readers. *NPR*. <https://www.npr.org/sections/parallels/2015/08/19/432779251/up-from-the-ashes-a-public-library-in-sri-lanka-welcomes-new-readers> [perma.cc/45ZX-CS3G]
- Meta. (2022). *How we create and use market-specific slur lists*. <https://transparency.meta.com/en-gb/enforcement/taking-action/how-we-create-and-use-market-slurs/> [perma.cc/9VN6-NUDU]




- Mirza, R. (2023, November 28). Case Study: Integrity or Influence? Facebook's Governance Trade-offs in India and the Power of the Press. *Shorenstein Center*. <https://shorensteincenter.org/case-study-integrity-influence-facebooks-governance-trade-offs-india-power-press/> [perma.cc/P74B-26JE]
- Mukherjee, R., & Desai, S. (2024). *Anatomy of a "Chapri": Exploring the Decontextualization of Casteist Slurs on Social Media*. [https://drive.google.com/file/d/1k6Cos64FGuQ1uXSEbf583TRyp\\_nawWKc/view?usp=embed\\_facebook](https://drive.google.com/file/d/1k6Cos64FGuQ1uXSEbf583TRyp_nawWKc/view?usp=embed_facebook) [perma.cc/Z2BR-7AZ5]
- Murugan, B., & Visalakshi, P. (2024). Ancient Tamil inscription recognition using detect, recognize and labelling, interpreter framework of text method. *Heritage Science*, 12(1), 1–21. <https://doi.org/10.1186/s40494-024-01522-9> [perma.cc/3E2P-QEKN]
- Nadaradjane, A. (2022, November 28). Sri Lanka: Democracy in Crisis. *The Centre for Independent Studies*. <https://www.cis.org.au/publication/sri-lanka-democracy-in-crisis/> [perma.cc/E55M-VY GK]
- Nanmalar, M., Vijayalakshmi, P., & Nagarajan, T. (2022). Literary and Colloquial Tamil Dialect Identification. *Circuits, Systems, and Signal Processing*, 41(7), 4004–4027. <https://doi.org/10.1007/s00034-022-01971-2> [perma.cc/QHL9-A4A2]
- Nicholas, G. (2022). *Shedding Light on Shadowbanning*. Center For Democracy And Technology. <https://cdt.org/insights/shedding-light-on-shadowbanning/> [perma.cc/8XNP-UDFS]
- Nicholas, G., & Bhatia, A. (2023a). *Lost in Translation: Large Language Models in Non-English Content Analysis* (arXiv:2306.07377). arXiv. <https://doi.org/10.48550/arXiv.2306.07377> [perma.cc/RWL3-S3DM]
- Nicholas, G., & Bhatia, A. (2023b). Toward Better Automated Content Moderation in Low-Resource Languages. *Journal of Online Trust and Safety*, 2(1), Article 1. <https://doi.org/10.54501/jots.v2i1.150> [perma.cc/KQ87-GRVG]
- Noolaham Foundation. (n.d.). Noolaham Foundation Library. Retrieved April 5, 2025, from [https://noolaham.org/wiki/index.php/முதற்\\_பக்கம்](https://noolaham.org/wiki/index.php/முதற்_பக்கம்) [perma.cc/9B4E-CBSA]
- Ramesh, A., Parthasarathy, V. B., Haque, R., & Way, A. (2020, December 4). *An error-based investigation of statistical and neural machine translation performance on Hindi-to-Tamil and English-to-Tamil*. 7th Workshop on Asian Translation (WAT2020), Suzhou, China (Online). <https://www.aclweb.org/anthology/2020.wat-1.22> [perma.cc/B6J5-TE98]
- Rowe, J. (2022). Marginalised languages and the content moderation challenge. *Global Partners Digital*.
- Savolainen, L. (2022). The shadow banning controversy: Perceived governance and algorithmic folklore. *Media, Culture & Society*, 44(6), 1091–1109. <https://doi.org/10.1177/01634437221077174> [perma.cc/AH5B-YHD9]
- Schiffman, H. (1996). *Linguistic Culture and Language Policy*. <https://www.sas.upenn.edu/~haroldfs/sars238/tamil238.html> [perma.cc/U56P-RZ89]


- Sehgal, D., & Nambiar, L. T. (2024). *Online Gender Based Violence on Short Form Video Platforms*. Centre for Internet and Society. <https://cis-india.org/raw/online-gender-based-violence-on-short-form-video-platforms> [perma.cc/9GD8-C538]
- Shahid, F., & Vashistha, A. (2023). Decolonizing Content Moderation: Does Uniform Global Community Standard Resemble Utopian Equality or Western Power Hegemony? *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–18. <https://doi.org/10.1145/3544548.3581538> [perma.cc/36U6-EMBF]
- Sivapriyan, E. (2025). Explained | Tamil Nadu's 2-language policy and opposition to NEP & Hindi. *Deccan Herald*. <https://www.deccanherald.com//india/tamil-nadu/explained-tamil-nadus-2-language-policy-and-opposition-to-nep-hindi-3422150> [perma.cc/Y6KD-TH8Y]
- Sombatpoonsiri, J., & Mahapatra, S. (2024). *Regulation or Repression? Government Influence on Political Content Moderation in India and Thailand*. Carnegie Endowment for International Peace. <https://carnegieendowment.org/research/2024/07/india-thailand-social-media-moderation?lang=en> [perma.cc/8L6S-UBXU]
- Soundararajan, T., Kumar, A., Nair, P., & Greely, J. (2019). *Facebook India: Towards the Tipping point of voilence*. Equality Labs. [https://equalitylabs.wpengine.com/wp-content/uploads/2023/10/Facebook\\_India\\_Report\\_Equality\\_Labs.pdf](https://equalitylabs.wpengine.com/wp-content/uploads/2023/10/Facebook_India_Report_Equality_Labs.pdf) [perma.cc/D2WJ-CJUP]
- Subramaniam, S. (n.d.). Should A Country Speak a Single Language? *The New Yorker*. Retrieved February 24, 2025, from <https://www.newyorker.com/magazine/2024/11/25/should-a-country-speak-a-single-language> [perma.cc/XRS5-MY8F]
- Tamil Guardian. (2023). *Tamil Eelam flags fly high in Canada and UK*. <https://www.tamilguardian.com/content/tamil-eelam-flags-fly-high-canada-and-uk> [perma.cc/Q5PQ-SQ7D]
- Tamil Guardian. (2024). Sri Lanka arrests Tamil man under PTA for Facebook post | Tamil Guardian. *Tamil Guardian*. <https://www.tamilguardian.com/content/sri-lanka-arrests-tamil-man-under-pta-facebook-post> [https://perma.cc/4YPA-L2J]
- Tattle. (2024). *Online Gender-Based Slurs and Abuses Don't Just Happen in English*. <https://tattle.co.in/blog/2024-05-09-slurs-occur-every-language/> [perma.cc/AA3A-GKJH]
- Thakur, D., & Liansó, E. (2021). Do You See What I See? *Capabilities and Limits of Automated Multimedia Content Analysis*. <https://cdt.org/insights/do-you-see-what-i-see-capabilities-and-limits-of-automated-multimedia-content-analysis/> [perma.cc/U4LX-2SY8]
- Thakur, D., & Hankerson, D. L. (2022). *An Unrepresentative Democracy: How Disinformation and Online Abuse Hinder Women of Color Political Candidates in the United States*. Center For Democracy And Technology. <https://cdt.org/insights/an-unrepresentative-democracy-how-disinformation-and-online-abuse-hinder-women-of-color-political-candidates-in-the-united-states/> [perma.cc/FM2L-8VCP]
- Tiku, N. (2022, June 2). Google's plan to talk about caste bias led to 'division and rancor.' *Washington Post*. <https://www.washingtonpost.com/technology/2022/06/02/google-caste-equality-labs-tanuja-gupta/> [perma.cc/5SSL-99ME]




- Times of India. (2024, September 13). DMK and BJP clash over Tamil Nadu's two-language policy amid NEP controversy. *The Times of India*. <https://timesofindia.indiatimes.com/city/chennai/dmk-and-bjp-clash-over-tamil-nadus-two-language-policy-amid-nep-controversy/articleshow/113321462.cms> [perma.cc/M2LW-CXQG]
- Treem, J. W., Leonardi, P. M., & van den Hooff, B. (2020). Computer-Mediated Communication in the Age of Communication Visibility. *Journal of Computer-Mediated Communication*, 25(1), 44–59. <https://doi.org/10.1093/jcmc/zmz024> [perma.cc/9QK4-T9BW]
- u/diegomrosa. (2023, July 20). *Did some Facebook languages disappeared? Which was the highest number of languages supported by Facebook?* [Reddit Post]. R/Facebook. [www.reddit.com/r/facebook/comments/154rmtt/did\\_some\\_facebook\\_languages\\_disappeared\\_which\\_was/](https://www.reddit.com/r/facebook/comments/154rmtt/did_some_facebook_languages_disappeared_which_was/) [perma.cc/DUH6-KLCQ]
- UN. (2011). *UN Guiding Principles*. Business & Human Rights Resource Centre. <https://www.business-humanrights.org/en/big-issues/governing-business-human-rights/un-guiding-principles/> [perma.cc/ZP2V-HQZ3]
- UNHCR IRIN. (2012). *Bridging the language divide in Sri Lanka*. [https://web.archive.archive.unhcr.org/20230527001057oe\\_/https://www.refworld.org/docid/501005892.html](https://web.archive.archive.unhcr.org/20230527001057oe_/https://www.refworld.org/docid/501005892.html) [perma.cc/4YVC-MWNG]
- Venkatachalapathy, A. R. (2022). *The Brief History of a Very Big Book: The Making of the Tamil Encyclopaedia*. Permanent Black.
- Vesteinsson, K. (2024). *Ahead of Landmark Elections, India's Government Silences Dissent*. Freedom House. <https://freedomhouse.org/article/ahead-landmark-elections-indias-government-silences-dissent> [perma.cc/RPU5-99KN]
- Vilk, V. (2023, June 29). *Shouting into the Void*. PEN America. <https://pen.org/report/shouting-into-the-void/> [perma.cc/J2NL-ZD37]
- Watts, I., Gumma, V., Yadavalli, A., Seshadri, V., Manohar, S., & Sitaram, S. (2024). *PARIKSHA: A Scalable, Democratic, Transparent Evaluation Platform for Assessing Indic Large Language Models*. <https://www.microsoft.com/en-us/research/publication/pariksha-a-scalable-democratic-transparent-evaluation-platform-for-assessing-indic-large-language-models/> [perma.cc/3PSQ-9UD3]
- Yatharth. (2025). *Bahujan Digital Publishing Infrastructures*. Centre for Internet and Society. <https://cis-india.org/raw/bahujan-digital-publishing-infrastructures> [perma.cc/Q2WF-ANE2]

 [cdt.org](https://cdt.org)

 [cdt.org/contact](https://cdt.org/contact)

 **Center for Democracy & Technology**  
1401 K Street NW, Suite 200  
Washington, D.C. 20005

 202-637-9800

 @CenDemTech

