



A report from



Research

Content Moderation in the Global South: A Comparative Study of Four Low-Resource Languages

Mona Elswah
Aliya Bhatia
Dhanaraj Thakur

June 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.

MONA ELSWAH
ALIYA BHATIA
DHANARAJ THAKUR

Authors



Content Moderation in the Global South: A Comparative Study of Four Low-Resource Languages

CDT Research Report

Mona Elswah, Aliya Bhatia, and Dhanaraj Thakur

WITH CONTRIBUTIONS BY

Samir Jain and DeVan L. Hankerson.

Illustrations and additional design by Osheen Siva.

ACKNOWLEDGMENTS

We would like to thank our local partners for their invaluable support during the project: Digital Citizenship, Paradigm Initiative, Center for Internet and Society, Hiperderecho. We thank the members of our advisory committee: Afef Abrougui, Juan Carlos Lara, Nanjira Sambuli and Jillian York. We also thank the study participants who generously shared their time and insights with us. We also gratefully acknowledge the feedback and suggestions of Kate Ruane and Michal Luria. Lastly, this work could not have been completed without the support of DeVan L. Hankerson, Timothy Hoagland, Drew Courtney, George Slover, Kevin Donnelly and Sarah Zolad. All views in this report are those of CDT.

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

Suggested Citation: Elswah, M., Bhatia, A., & Thakur, D., (2025). Content Moderation in the Global South: A Comparative Study of Four Low-Resource Languages. Center for Democracy & Technology. <https://cdt.org/insights/content-moderation-in-the-global-south-a-comparative-study-of-four-low-resource-languages/>

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.



Contents

Executive Summary: Insights from Four Case Studies	5
Background	8
1. The Global South Content Moderation Project	8
2. Content moderation: Realities and Knowledge Gaps	10
3. A World of Low-Resource Languages	12
Comparative Findings	14
1. Content moderation experiences in the Global South	14
A. Use of platforms	14
B. Widespread experience with content moderation errors	16
2. Companies employed multiple approaches when developing and enforcing policies: Global v. Local Approaches	20
A. Global Content Moderation	20
B. Local Content Moderation Approach	23
C. Alternative Paths: One Language, Multi-Country Approach	25
3. Manual and Automated Approaches to Enforcing Policy	26
4. Users experienced moderation differently and inconsistently	33
5. Tactics of Resistance & Fighting Back	33
Recommendations	37
Appendix	41
Methods & Data Collection	41
References	43

Executive Summary: Insights from Four Case Studies



Over the past 18 months, the Center for Democracy & Technology (CDT) has been studying how content moderation systems operate across multiple regions in the Global South, with a focus on South Asia, North and East Africa, and South America. Our team studied four languages: the different Maghrebi Arabic Dialects (Elsawah, 2024a), Kiswahili (Elsawah, 2024b), Tamil (Bhatia & Elswah, 2025), and Quechua (Thakur, 2025). These languages and dialects are considered “low resource” due to the scarcity of training data available to develop equitable and accurate AI models for them. To study content moderation in these languages spoken predominantly in the Global South, we interviewed social media users, digital rights advocates, language activists, representatives from tech companies, content moderators, and creators. We distributed an online survey to 562 frequent social media users across multiple regions in the Global South. We organized roundtables, focus group sessions, and talks to get to know these regions and the content moderation challenges they often face. We did this through essential collaborations with regional civil society organizations in the Global South to help us understand the local dynamics of their digital environments.

When we initially delved into this topic, we recognized that the culture of secrecy that surrounds content moderation would pose challenges in our investigation. Content moderation remains an area that technology companies keep largely inaccessible to public scrutiny, except for the information they choose to disclose. It is a field where the majority, if not all, participants are discouraged from engaging in external studies like this or revealing the specifics of their operations. Despite this, we gathered invaluable data and accessed communities that had previously not been reached. Our findings significantly contribute to the scientific and policy communities’ understanding of content moderation and its challenges in the Global South. The data we present in this report also contributes to our understanding of the information environment in the Global South, which is understudied in current scholarship.

Here, we compare and synthesize the insights we gained from studying the four regions and present our recommendations for improving content moderation in low-resource languages of the Global South. While the insights from this project may be applicable to other non-Western contexts and low-resource or indigenous languages, we have learned that each language carries its own rich history and linguistic uniqueness, which must be acknowledged when discussing content moderation in general. By comparing these four case studies, we can identify some of the overall content moderation challenges that face languages in the Global South. Additionally, this comparison can help

us identify the particular challenges inherent in moderating diverse linguistic and cultural contexts, enhancing our understanding of what could possibly be “effective” content moderation for these regions and beyond.

While we acknowledge the uniqueness of each language, when comparing the four languages we examined, we find that:

1. The content moderation policies currently employed by large tech companies have limitations. **Currently, global tech companies use two main approaches to content moderation: Global and Local. The global approach involves applying a uniform set of policies to all users worldwide. While this approach helps prevent external interventions (e.g., by governments) and is in some ways easier, it ignores unique linguistic and cultural nuances. The local approach, exemplified by TikTok, involves tailoring policies, particularly those related to cultural matters, to specific regions. This approach, despite its promise of inclusivity, sometimes poses obstacles and limitations on users trying to challenge local norms that violate their rights.** An exception to the two approaches was found in the Kiswahili case: JamiiForums, a Tanzanian platform, has developed its own unique methods for moderating local languages, introducing what we referred to as “multi-country approach.” Their unique approach, which entails assigning moderators to content from their native language, poses more promise and large user satisfaction, but leaves a question of whether it can be applicable on a large scale.

2. **Users in the Global South are increasingly concerned about the spread of misinformation and hate speech on social media** in their regions. All four case studies highlighted user concerns regarding the spread of hate speech and harassment and inconsistent moderation of the same. **Additionally, users are increasingly worried about the wrongful removal of their content,** particularly in the Tamil and Quechua cases. **Tamil and Quechua users linked the content restrictions to the companies’ desire to “silence their voices” more often than Kiswahili and Maghrebi Arabic-speaking users.**

3. **We identified four major outsourcing service providers that dominate the content moderation market for the low-resource languages we examined:** Teleperformance, Majorel, Sama, and Concentrix.

4. Across the four cases, we found that content moderators for non-English languages are often exploited, overworked, and underpaid. They endure emotional turmoil from reviewing disturbing content for long hours, with minimal psychological support and few wellbeing breaks. Additionally, we found that the hiring process for moderators lacks diversity and cultural competencies.

5. Moderators from a single country are often tasked with moderating content from across their region, despite dialectical and contextual variations. In general, moderators are required to review content in dialects other than their own, which leads to many moderation errors. In some cases, moderators are assigned English-language content from around the world, with no regard for their familiarity with specific regional contexts, as long as they possess a basic understanding of English.

6. Resistance is a common phenomenon among users in the Global South. Many users across the case studies employed various tactics to circumvent and even resist against what they saw as undue moderation. Despite the constant marginalization of their content and their languages, users developed various tactics to evade the algorithms, commonly known as “algospeak”. We found tactics that involved changing letters in the language, using emojis, uploading random content alongside material they believed would be restricted, and avoiding certain words. In examples from our Quechua case study, some simply posted in Quechua (instead of Spanish) because they found that it was often unmoderated.

7. Lastly, many NLP researchers and language technology experts in the Global South have developed tools and strategies to improve moderation in many low-resource languages. They have engaged with their local communities to collect datasets that represent specific dialects of the language. They enlisted students and friends to help annotate data and have published their work, creating networks to represent their language in global scholarship. However, these scholars and experts often feel underutilized or unheard by tech companies. If consulted and their knowledge utilized, these groups could significantly improve the current state of content moderation for low-resource languages.



Background



1. The Global South Content Moderation Project

In late 2023, the Center for Democracy & Technology (CDT) launched a research project, supported by a grant from the Internet Society Foundation, to investigate content moderation systems in the Global South. The purpose of this project was to build upon previous research conducted by CDT which highlighted the shortcomings of content moderation in non-English and non-Western contexts and to advance understanding of how social media content moderation works for languages in the Global South.

In this project, we focused on four languages: Maghrebi Arabic, Kiswahili, Tamil, and Quechua. Our selection was driven by several key factors. First, these languages represent distinct regions within the Global South, each with its unique geopolitical and cultural dynamics. Second, each language is spoken by a significant population, with estimates ranging from 10 million to 100 million speakers (see Table 1). Third, each language has been influenced by colonization in one way or another and has been the subject of concerted efforts to increase its use, proliferation, and preservation either online or off. Some speakers have developed new forms of the language, while others have integrated the dominant or colonially imposed language (e.g., English or Spanish) into their original languages. Lastly, these languages are classified as low-resource languages that lack high quality datasets needed for training AI models, which presents unique challenges in the arena of automated content moderation ([Shahid et al., 2025](#); [Nicholas & Bhatia, 2023](#)).

Our first case study examined Maghrebi Arabic dialects in North Africa, representing more than 100 million Arab speakers. We focused on the dialects spoken in Morocco, Tunisia, and Algeria for their socio-historical and cultural similarities, along with common linguistic characteristics ([Harrat et al., 2018](#)). Maghrebi Arabic dialects are colloquial forms of Arabic, with each dialect possessing its own distinct vocabulary and nuances. Certain words and phrases may carry different meanings or connotations across the Maghreb region ([Elsawah, 2024a](#)).

Our second case study focused on Kiswahili (also known as Swahili) in East and Central Africa ([Elsawah, 2024b](#)). More than 80 million people are estimated to speak Kiswahili across Africa. While Kiswahili is spoken in many countries, we focused our study on the Kiswahili variations spoken in Kenya and Tanzania. We selected Tanzania for having the largest Kiswahili-speaking population and for being the birthplace of standard Swahili. We selected Kenya due to the high number of

	Maghrebi Arabic	Kiswahili	Tamil	Quechua
Estimated Number of Speakers	~ 100 million	~ 80 million	~ 80 million	~ 10 million
Countries Examined in the Study	Morocco, Algeria, Tunisia	Kenya and Tanzania	India and Sri Lanka	Peru, Ecuador and Bolivia
Region	North Africa	East and Central Africa	South Asia	South America

▲ **Table 1.** A comparison between the four languages we included in the analysis.

Kiswahili speakers, second only to Tanzania, and for being Africa’s Silicon Savannah ([Dzahene-Quarshie, 2009](#); [Mwaura, 2023](#); [Wahome, 2023](#)).

We selected Tamil for our third case study ([Bhatia & Elswah, 2025](#)). Tamil is spoken by more than 80 million people worldwide ([Murugan & Visalakshi, 2024](#)) and is the 17th most spoken language in the world. Despite that, it is considered a low-resource language, in part due to the languages’ regional variation and history of politicization, particularly in Sri Lanka. Many Natural Language Processing (NLP) researchers, language technology experts and even a state government in India, specifically the Tamil Nadu government in southern India, have invested in efforts to digitize the language. We focused on Tamil speech and moderation experiences in India and Sri Lanka, as these are the two countries with the largest Tamil-speaking populations.

Lastly, we studied the Quechua language, one of the largest groups of indigenous languages in South America. This group of languages has around 10 million speakers and has evolved as a consequence of colonization from the European invasion of the continent in the 16th century. We focus on three Quechua language varieties from the Andean region: Central Quechua (Wanka, Ancash, Pataz), Southern Quechua (Chanka, Collao), and Quichua/Kichwa (Ecuador, Amazon). This covers some of the main varieties spoken in Peru, Ecuador, and Bolivia, where Quechua holds official status.

Members of our team have lived in some of these regions for many years and thus possess considerable knowledge of the cultural, digital, and political dynamics. However, as our team does not speak all of these languages, we decided that it was essential to collaborate with regional partners who were either native speakers of these languages or had access to native speakers. Therefore, we collaborated with Digital Citizenship in Tunisia, Center for Internet and Society in India, Hiperderecho in Peru, and Paradigm Initiative in Kenya. These collaborations helped us

to avoid making false assumptions or introducing unintentional bias to our study. Our regional partners reviewed our work and connected us with linguistic activists, content creators, advocates, moderators, and other key stakeholders. They served as our primary point of contact whenever we had questions about the languages or cultural nuances.

2. Content moderation: Realities and Knowledge Gaps

Content moderation determines the rules of participation within online communities and shapes what is seen and valued in digital spaces ([Grimmelman, 2015](#)). All platforms employ some form of content moderation of user-generated content ([Gillespie, 2018](#)). It is a “multi-dimensional process through which content produced by users is monitored, filtered, ordered, enhanced, monetised or deleted on social media platforms” ([Badouard & Bellon, 2025](#)). It not only includes analysis of user generated content, but also how that content is evaluated, how policies are enforced, how content moderation decisions are appealed, and how users are educated about such policies ([Kamara, et al. 2021](#)).

Content moderation strives to ensure a positive user experience, safeguard social media platforms’ reputations, enable compliance with legal requirements, and drive advertising revenues ([Roberts, 2019](#); [Barnes, 2022](#); [Gillespie, 2018](#)). Platforms are aware that too much divisive and harmful content will hurt their business and that they must keep “objectionable” content away from their platforms ([Barnes, 2022](#)). Thus, content moderation has become essential to maintaining the balance of profitability and user safety.

Nevertheless, in many instances, platforms have failed to moderate content appropriately and consistently, affecting users differently and at times leading to real world violence, like in the Myanmar genocide, the Israel-Palestine conflict ([Alimardani & Elswah, 2021](#); [Elswah, 2024a](#)), the Tigray genocide ([Crystal, 2023](#)), as well as to the proliferation of hate speech content in Global South countries like Malaysia ([Jalli, 2023](#)), Indonesia ([Renaldi, 2024](#)), Zimbabwe ([France 24, 2023](#)), Philippines ([Etter, 2017](#)), Nigeria ([Egbunike, 2020](#)), and many others. On the other hand, there are numerous instances where platforms have wrongly removed essential and harmless content and restricted journalist accounts and pages and hampered the ability for users to mobilize and document human rights abuses ([Nandakumar & Amarasingam, 2021](#)).

In response to increasing public and official scrutiny of their content moderation practices, platforms have been compelled to implement more robust mechanisms to address these concerns. Tech companies employ thousands of moderators worldwide, often through third-party companies, to review an endless stream of toxic content on a daily basis (Barnes, 2022; [Arsht & Etcovitch, 2018](#)). Moderators spend their days working in a high-pressure environment trying to meet strict numerical quotas while reviewing traumatic content at a rapid pace ([Arsht & Etcovitch, 2018](#)).

Additionally, companies have introduced automated content moderation systems and processes, including AI-powered tools, to optimize content moderation ([Gillespie, 2020](#)). While some in Silicon Valley envision a robust automated moderation system or “techno-solutionism” that replaces the controversial human element in content moderation (Barnes, 2022), evidence thus far shows that human moderation remains essential ([Gillespie, 2020](#)). The limitations of machine learning tools used for content analysis of user generated content (e.g., text, video, audio), have been well documented ([Shenkman, et al. 2021](#); [Duarte et al., 2017](#)).

Multi-stakeholderism has become a central aspect of discussions on content moderation, with many non-platform actors attempting to define desired rules of moderation ([Badouard & Bellon, 2025](#); [Sombatpoonsiri & Mahapatra, 2024](#)). While the role of civil society and other actors in pressuring tech companies to modify and improve their content moderation systems — especially in the Global South — is often publicly acknowledged, the extent to which these stakeholders influence platforms and tech companies remains an open question. For many, civil society-led advocacy is often the only way users can gain clarity on and remedy moderation decisions ([Elswah, 2024a & Elswah, 2024b](#)). Furthermore, while tech companies release annual reports detailing government requests for content removals and data sharing, the precise, often hidden, interventions by governments in shaping companies’ policies are not transparent, particularly in countries where there are often restrictions on civil society and weaker rule of law ([Sombatpoonsiri & Mahapatra, 2024](#)).

There are a lot of unknowns in the field of content moderation, and much of our current understanding is focused on the Western world. This project, however, aims to shed light on specific aspects of content moderation in languages of the Global South.

3. A World of Low-Resource Languages

English is usually considered the official language of the Internet ([Sengupta et al., 2020](#); [Wolk, 2004](#)). It is the most dominant language on the Internet, accounting for about 63.7% of websites ([Nicholas & Bhatia, 2023](#)) despite being spoken by only 16% of the world. Due to the historical and colonial prioritization of knowledge production in English and other Latin languages, often at the expense of production in indigenous languages, more written products and artifacts are in English than in any other language ([Nicholas & Bhatia, 2023](#)). US-based tech companies have further perpetuated this dominance of the English language online by making their digital services available primarily in English to global users without equal or adequate investment in considered low-resource languages ([Joshi et. al, 2020](#); [Robinson et al., 2023](#)).

A “low-resource language” is one that lacks high quality and diverse training datasets representing the language, despite the number of people who speak or write in this language. Many indigenous and endangered languages are considered low-resource. Often, a measure of resourcedness correlates with digital access and digital penetration, with many communities who have faced barriers in accessing digital devices and hardware and software in their languages having a lower likelihood of creating online content in those same languages ([Mahmud et al., 2023](#)). It can also include the less studied and less computerized languages ([Mahmud et al., 2023](#)). Additionally, the resourcedness of a language is not always correlated with the number of speakers ([Nicholas & Bhatia, 2023](#)), but instead on prioritization made by tech companies and perception of users’ value towards their bottom line.

Digitizing and creating tools in multiple languages requires a great deal of resources. Often, tech companies have believed that users speaking low-resource languages are less economically valuable than users in the Global North. They’ve also implicitly assumed that all languages and communities around the world subscribe to Anglocentric values embedded in English, ultimately resulting in a deprioritization of languages and users in the Global South ([Sarveswaran, 2024](#)). Academic researchers have likened this to a form of “digital colonialism” where global digital platforms scale or prioritize English at the expense of other languages, replicating the asymmetries of power seen during colonial eras, particularly where low- and high-resource languages existed in relative proximity throughout the colonial period to the present day

([Shahid et al., 2025](#)). While this feature is present in all our four case studies (e.g., with English and French), this is particularly the case with Spanish and Quechua.

This gap in languages has further implications when AI is considered; English currently dominates the field of AI and Natural Language Processing, the specific sub-field of examining and building language technologies, due to the existence of vastly more digitized text data in English than any other language ([Joshi et. al, 2020](#)). This has made it easier to build large language models in English compared to any other language. Curating and labeling high-quality datasets of low-resource languages is one of the biggest challenges in the NLP field ([Mahmud et al., 2023](#)). Current generative AI systems, based specifically but not exclusively on large language models (LLMs), perform poorly in low-resource languages ([Robinson et al., 2023](#)). Non-English language data scraped from the Internet are usually machine translated or scanned from an image, a process that might include many errors ([Nicholas & Bhatia, 2023](#)).

Thus, in this report, we compare the ways low resource languages are impacted by the lack of resources and how NLP researchers and digital language experts deal with this. In the following sections, we present our comparative findings and end with recommendations based on these insights.



Comparative Findings



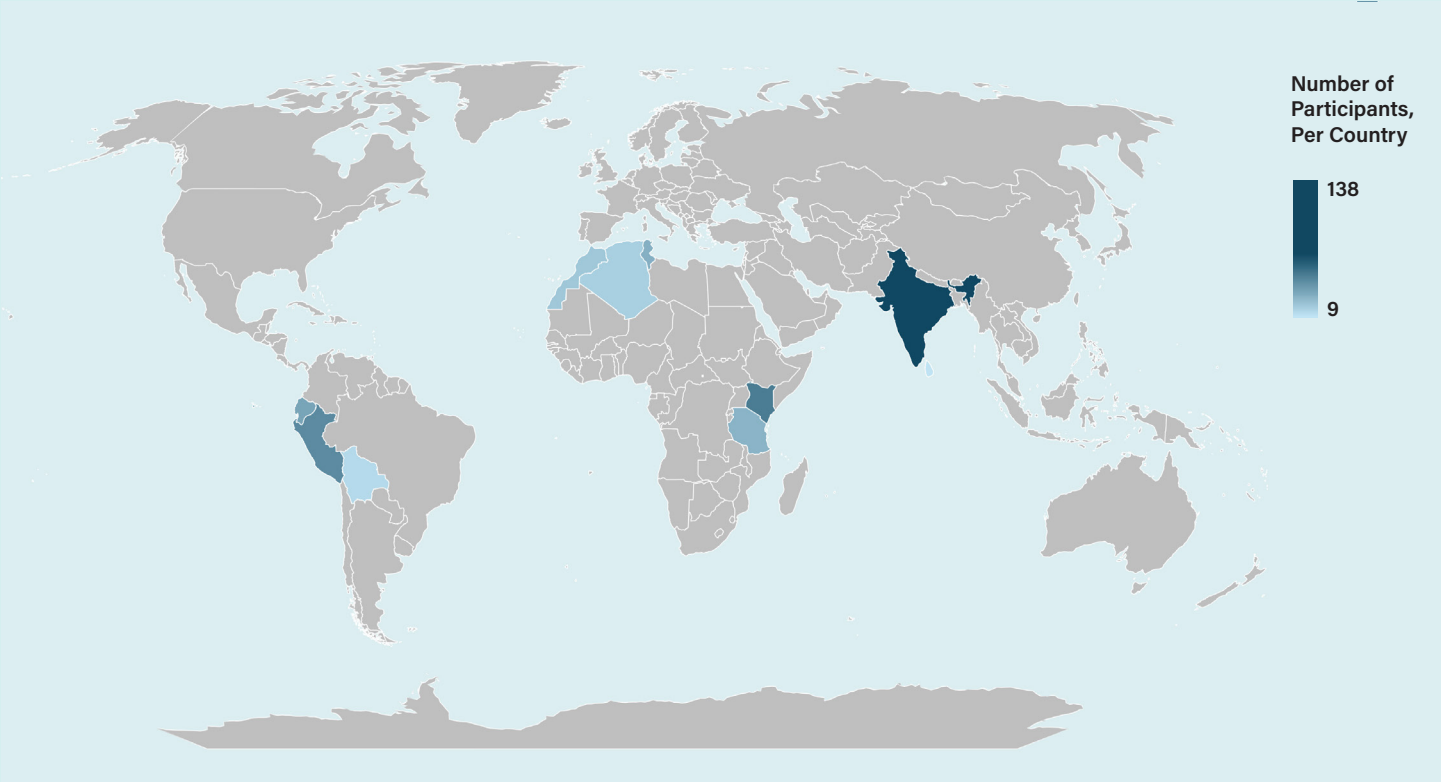
1. Content moderation experiences in the Global South

In our 18-month investigation, we surveyed frequent social media users across the four regions we examined and gained many insights into how users perceive content moderation in their language and contexts. We surveyed 562 respondents from Tunisia, Morocco, Algeria, Kenya, Tanzania, India, Sri Lanka, Peru, Ecuador, and Bolivia (See figure 1). The survey examined users' perceptions of content moderation, their trust of social media platforms, and their online experiences in relation to content removal and communication with the platforms. Overwhelmingly, users believe they experience inconsistent moderation, shadowbanning or opaque content moderation, or government-influenced moderation, particularly when they speak about certain topics including politics and harassment.

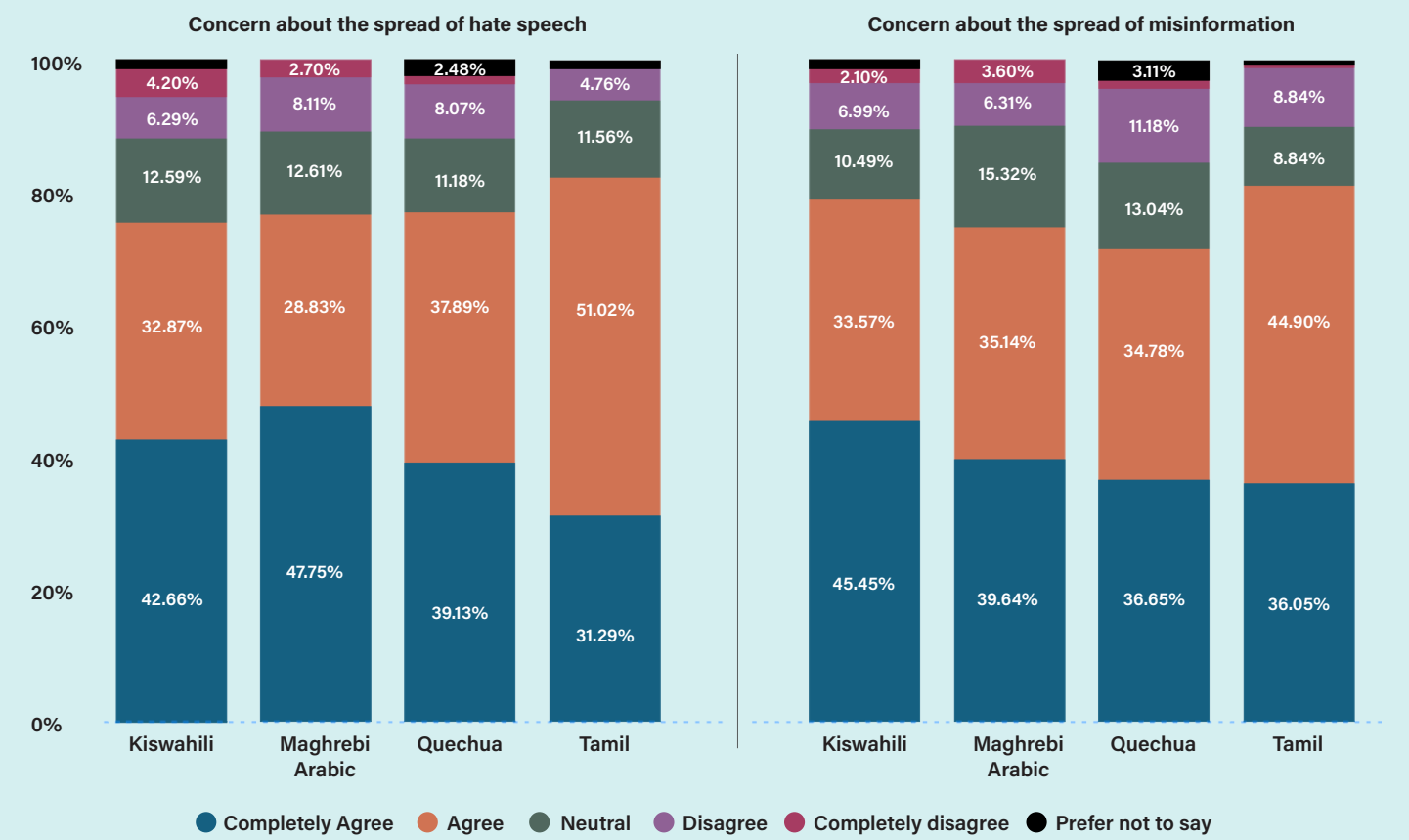
A. USE OF PLATFORMS

The survey conducted across languages found that the following platforms were most used by respondents in our four cases: Facebook, YouTube, and Instagram. X remained popular across all regions except the Maghreb, where TikTok and LinkedIn showed higher usage. TikTok's popularity is rapidly growing in the Global South — although not in regions such as in India where the service is banned. Nevertheless, TikTok use has surpassed X in both the Quechua and Maghrebi Arabic contexts. In the aftermath of India's 2020 TikTok ban, a number of other video-sharing mobile apps became prominent, such as Chingari, which 60% of Tamil respondents stated they used Moj, Josh, and MX Takatak.

According to our participants, misinformation and hate speech are big concerns in the information ecosystems. Tamil users were the most concerned about both types of harmful content in their region, with 82.31% expressing concern about hate speech (51.02% agreeing and 31.29% completely agreeing), and 80.95% concerned about misinformation (44.90% agreeing and 36.05% completely agreeing) (See Figure 2). Also, Kiswahili users showed a high level of concern about misinformation, with 79.02% expressing concern (45.45% completely agreeing and 33.57% agreeing), and 75.53% concerned about the spread of hate speech. Maghrebi Arabic users followed closely, with about 77% either agreeing or completely agreeing about the concern with hate speech and 75% worried about misinformation (39.64% completely



▲ **Figure 1.** Countries included in the survey data (n=562). Source: CDT’s online survey (April 2024-March 2025).



▲ **Figure 2.** Users’ concerns about the spread of hate speech and misinformation across the four case studies (n=562). Source: CDT’s online survey (April 2024- March 2025).

agreeing and 35.14% agreeing). Quechua users also reported relatively high levels of worry about both misinformation and hate speech, with about 71.3% and 76.9% respectively either agreeing or completely agreeing.

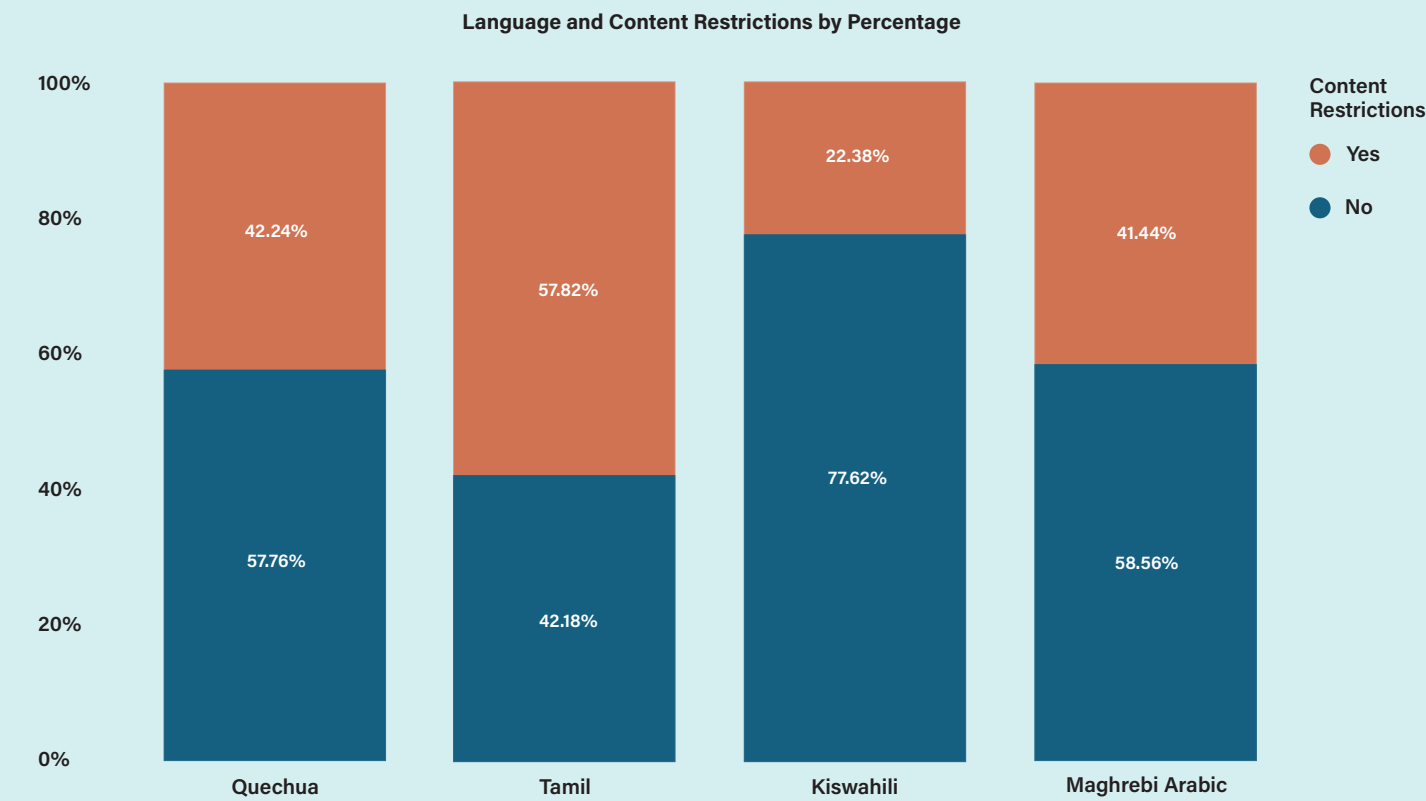
In response to the spread of misinformation and hate speech in their regions, participants sought to report harmful content themselves. They often viewed this as an altruistic act or a way to control their online experience. Approximately 67% of participants across the four regions indicated that they had reported content they perceived as violating platform guidelines. While the majority of users in our study have reported at least one incident on a social media platform, users felt kept in the dark about platform action in response to their reporting. Among the four cases, Tamil speakers were the most likely to note that social media platforms ignored their reports. In contrast, Kiswahili participants reported more positive outcomes from their reporting experience compared to participants from other language groups.

B. WIDESPREAD EXPERIENCE WITH CONTENT MODERATION ERRORS

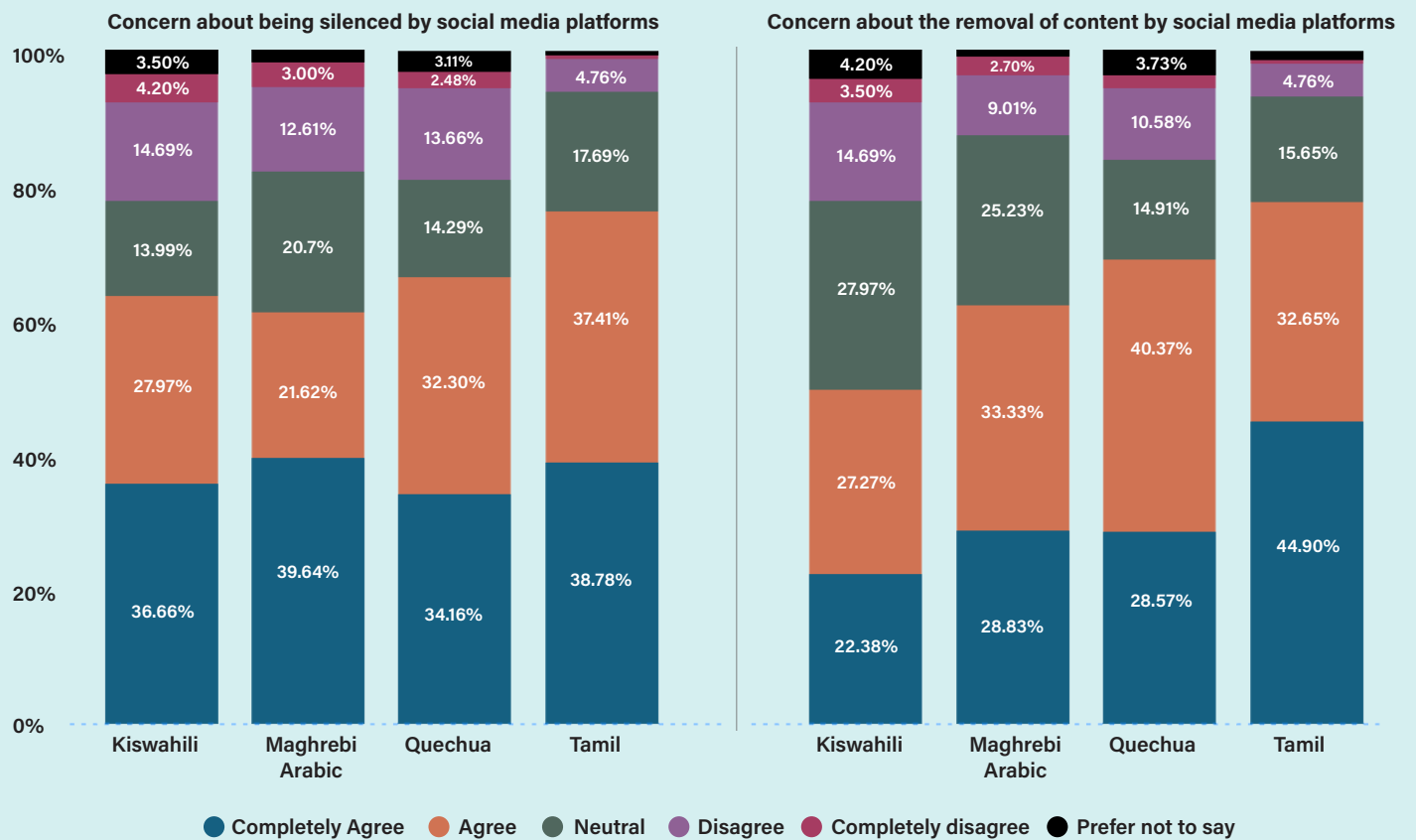
While survey respondents believed violating content was not always removed, a significant portion (41%) of survey respondents reported facing content restrictions or removals themselves. Across the four case studies, survey respondents reported the highest rates of content removals in the Tamil case study with 57.8% of respondents noting that they had faced removal of content in the past, as compared to 41.4% in the Maghrebi Arabic study (See Figure 3). In one interview, a user who posted online in Tamil noted that they had read the terms of service multiple times, yet still faced removal even when they believed their content did not violate the terms of service.

The second largest group of participants experiencing content suppression was in the Quechua context. This experience was often linked to the belief of widespread linguistic suppression of the Quechua language and the historical marginalization of its speakers. Lastly, Kiswahili had the lowest incidence of content removals, which aligns with the qualitative data we collected, suggesting that Kiswahili content faced fewer restrictions compared to the other languages.

Overall, a significant percentage across the four groups expressed concern about content removal. Tamil participants showed the highest level of concern about content restrictions on social media, which aligns with the previous finding that they experienced the highest level of content restrictions and removals — about 77.5% either agreed or



▲ Figure 3. Percentage of participants who reported experiencing content removal or restrictions (n=562). Source: CDT’s online survey (April 2024- March 2025).

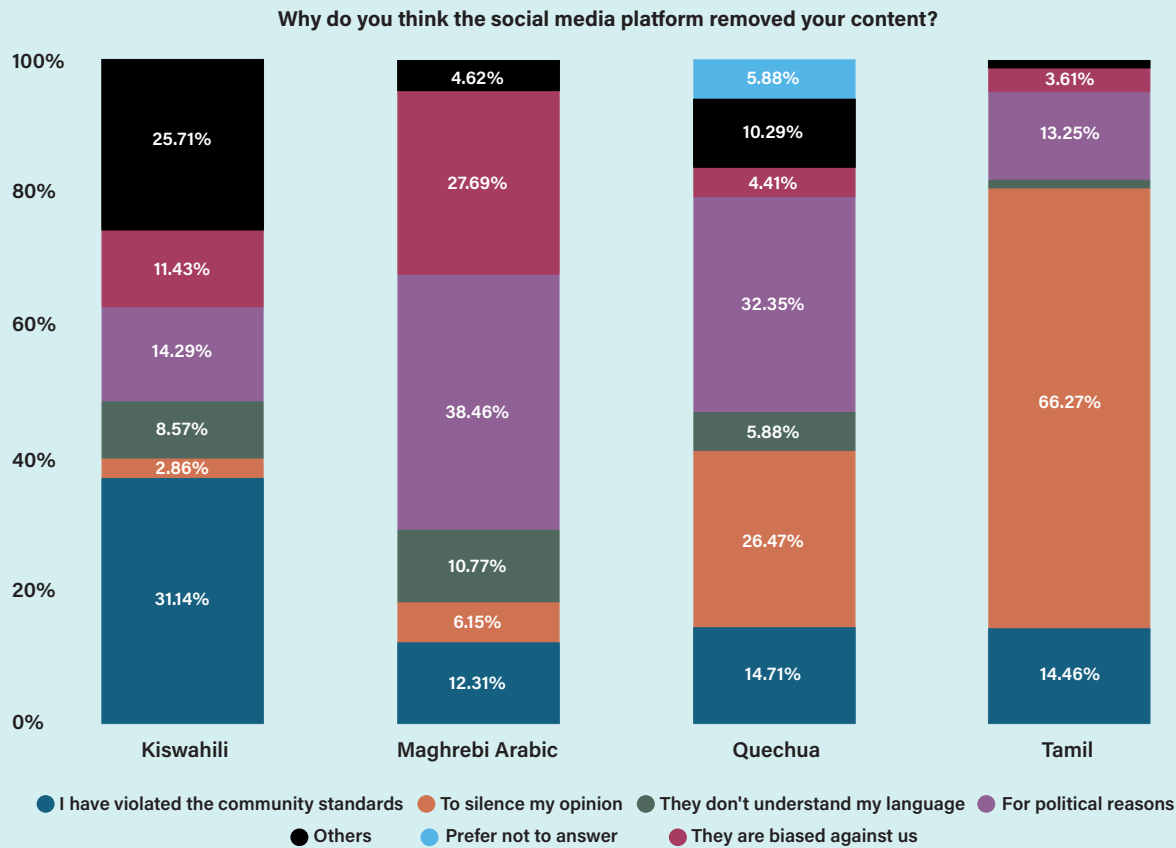


▲ Figure 4. Concerns about being silenced by social media and the removal of content by social media (n=562). Source: CDT’s online survey (April 2024- March 2025).

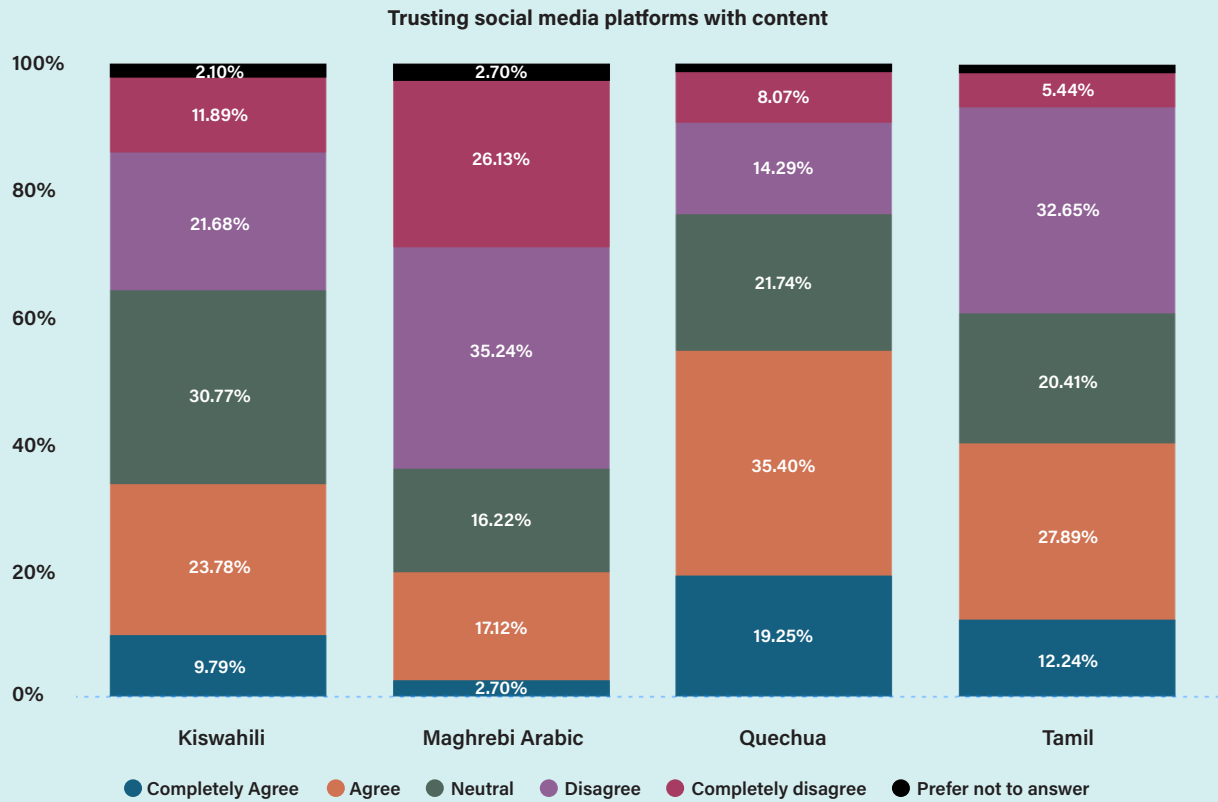
completely agreed that they were concerned with content removals. Similarly, the Quechua community was the second most concerned, with about 69%, followed by 62% of Maghrebi Arabic participants (See Figure 4). The Kiswahili group, while the least concerned with 49.6%, still had a large portion of participants expressing worries. These results indicate that content removal is a significant issue for users in the Global South, particularly in the Tamil context.

In light of these restrictions, we also explored participants' concerns about being silenced by social media companies, which closely align with the findings on content removals. Tamil participants, once again, showed the highest level of concern, with about 76% either agreeing or completely agreeing that they fear being silenced on social media (See Figure 4). Similarly, the Quechua community, which ranked second in terms of content removals expressed similar concerns about removals (69% either completely agreeing or agreeing) and also reported significant concern about being silenced (66.4% either completely agreeing or agreeing). This aligns with both the Tamil and Quechua communities' experiences of linguistic suppression, further highlighting the vulnerability felt by these groups. Kiswahili and Maghrebi Arabic users surveyed showed significant levels of concern about being silenced by social media, with about 64.5% and 61% respectively completely agreeing and agreeing, indicating heightened sensitivity to censorship, although less pronounced than in Tamil and Quechua groups. These results suggest that concerns about content removal and being silenced are intertwined, with Tamil and Quechua groups experiencing the greatest fear of censorship.

Across all four regions, the perception of censorship outweighs the other possible reasons for content removal. In our online survey, participants agreed with statements asserting that content removals or restrictions were attempts to: "silence their opinions" or "for political reasons" (See Figure 5). Tamil users, in particular, were at the top of the list of those who perceived content restrictions as attempts to silence their voices online, with approximately 66.7% reporting that these restrictions were likely attempts to suppress their voices. This can likely be attributed to a number of factors including increased legal authority to request removal of content in countries where Tamil is widely spoken, and the historic politicization of the North and South in India and the marginalization of the Tamil language in Sri Lanka driving many Tamil speakers to suspect politically motivated censorship. Maghrebi Arabic and Quechua users attributed "political reasons" as the main reason behind content restrictions and removals, with 38.4% and 32.3% respectively. During the interviews, Maghrebi Arabic



▲ **Figure 5.** Perceived reasons of content removal (% of participants who stated they had their content restricted or removed, n=251). Source: CDT’s online survey (April 2024- March 2025).



▲ **Figure 6.** Survey respondents’ self-reported level of trust towards social media companies (n=562). Source: CDT’s online survey (April 2024- March 2025).

users noted that these restrictions happened frequently with content related to the Gaza war. In our interviews, Quechua social media users suggested they sometimes posted about Quechua people's rights and related political issues in their respective countries and that this was subject to removal. By contrast, in the Kiswahili case, where fewer restrictions were reported by participants, survey participants cited “violating community standards” as a main reason for content removal.

Overall, participants tend to distrust social media platforms, particularly Maghrebi Arabic respondents, with about 61% of participants stating that they disagree and completely disagree about trusting social media platforms with their content (35.14% and 26.13% respectively) (See Figure 6). This is in line with our qualitative analysis showing that the many reported incidents of content restrictions and shadowbanning in relation to the Gaza war have impacted this trust, particularly with US-based social media companies. Tamil and Kiswahili participants showed high levels of distrust against social media platforms, with 38.1% and 33.6% respectively indicating either disagreement or strong disagreement. Quechua participants were the most trusting of social media platforms: more than 54% stated that they trust social media platforms to an extent (35.4% agree and 19.25% completely agree that they trust social media platforms) despite this community's experience of different moderation standards applied to Spanish than to Quechua. This could be a result of the differing experiences of respondents who post in Quechua (or both Spanish and Quechua) and those who post only in Spanish.

2. Companies employed multiple approaches when developing and enforcing policies: Global v. Local Approaches

How do tech international companies moderate user generated content from the Global South? Across the four case studies, we identified two main approaches.

A. GLOBAL CONTENT MODERATION

First, we found that US-based tech companies typically adopt what we refer to as a global content moderation approach. This approach involves applying the same policies to all users, regardless of location, language, culture, or religion, except in some specific cases of elections

or crises that necessitate regional, urgent, and temporary modifications. Many of these “global” policies applied uniformly across regions and users often originate from U.S. laws and norms. Others are informed loosely by international human rights instruments ([Dvoskin, 2023](#)).

During our interviews, representatives from tech companies say companies employ the global approach for multiple reasons. First, it enables application of policies at scale by applying a similar set of policies and principles across all the content they host, which often amounts to millions if not billions of posts per year. Second, platforms say that this enables equal treatment of posts no matter where the user is based, enabling the free flow of communication. Instead of having to apply a policy depending on region, a user in region A posting about a news event will be treated the same as a user in region B posting about the same news event. Additionally, platforms explained that their approach was designed to prevent potential government interventions and uphold human rights principles, particularly in instances where authoritarian regimes might seek to silence dissenting voices. Companies often defend this approach by claiming that their policies are uniform, leaving no room for “edits” by external actors (with exceptions for crisis events such as natural disasters, political coups, etc.). Whether this approach effectively prevents government or external intervention in content moderation remains a topic for further exploration.

Uniform or “global” applications of content policies also enable convenient oversight. When we interviewed content moderators, they shared that the approach was easier for them to understand, as they only needed to learn one policy and a consistent set of interpretations. Moreover, company policy staff can easily identify mistakes or make edits to one set of policies with the global approach.

However this model does not account for the unique linguistic variations and cultural and political dynamics in each country, resulting in numerous reports of bias and content moderation errors ([Elswah, 2024c](#)). Moderators also struggle to apply policies across a range of regions and political contexts, especially when they are not familiar with the Western context from which the policies originate.

For example, moderators in the Maghreb region mentioned that they were required to approve content featuring immodesty permitted under the policies of the US-based platform but not aligned with local cultural norms. In the case of Quechua, former content moderators described how they often encountered numerous problems when they tried to apply a content policy that had been translated from English to Spanish

and then to Quechua. This included words, phrases, or concepts meant to capture offensive speech that were not relevant in Quechua. Similarly, users in the Tamil case study noted that platforms did not moderate caste-based harassment, despite its similarity to harassment that platforms did not permit, simply because U.S.-based policies did not include caste as an example of a characteristic that could be used to systematically target individuals.

Additionally, across the four cases, we found that when hiring policy teams responsible for overseeing these regions companies did not always require or consider linguistic expertise in the respective languages. During our project, we interviewed several people responsible for shaping policies for different regions, many of whom lacked linguistic expertise for the regions they were developing policies for. In the Tamil case, a U.S.-based social media policy lead explained that they follow what they call a "coverage model," where the policy team would think about tailoring policy enforcement to a language only during times of crisis or when government actors applied pressure on a company to address certain events. When language did come up, teams would rely on machine translation tools, which are widely known to be error prone. In the case of Quechua, we learned that a major social media company brought in indigenous speaking moderators only a few years ago, and the move marked the start of applying existing global content policies for Quechua. It is unclear who was responsible for addressing Quechua content before that time or how moderation was done.

Under this approach, moderators were tasked with reviewing content in their assigned languages and were also expected to review content in other languages and from other countries. For instance, moderators who were recruited to assess content in Quechua also reviewed content in Spanish across the South American region that included different cultural contexts. In at least two other cases, English was the other relevant language. For example, in East Africa, moderators reported that in addition to reviewing Kiswahili content they had to review content from English-speaking countries, such as the US, the UK, and Canada. In the Tamil case study, we found that moderators were required to review content that had been machine-translated into English without being informed of the original language of the content, and that moderators in the region were recruited for their English-language proficiency. This was not an issue in the Maghrebi Arabic case, as moderators in this region are not typically expected to be fluent in English.

Approach	Global	Local	Multi-Country
Definition	Employ generally uniform policies across all users worldwide.	Tailor some policies to account for local and national cultural nuances.	Tailor policies to address national cultural nuances and laws for each country.
Companies	Meta, X, YouTube	TikTok, Reddit, Quora, Sharechat	JamiiForums
Relevant Languages	Maghrebi Arabic, Kiswahili, Tamil, and Quechua	Maghrebi Arabic, Kiswahili, and Tamil	Kiswahili
Trade offs	Reduces external and non-platform interventions.	More sensitive to cultural and linguistic variations.	Respects national laws, culture, and linguistic context for each country.
	Easier for moderators to follow.	Prevents users from challenging the local norms.	Easy for moderators to follow.
	Does not account for cultural and linguistic variations.	Harder for moderators to follow.	Harder to employ on a large scale.

▲ **Table 2.** Comparison between the three approaches to content moderation.

B. LOCAL CONTENT MODERATION APPROACH

In contrast, some companies adopt methods to localize content moderation or employ local content moderation approaches in part to comply with government requests and regulation or adapt to user needs or criticism. We found that some tech companies follow what we refer to as a “local content moderation” approach in our three cases on Maghrebi Arabic, Kiswahili, and Tamil (See Table 2). Based on our interview data, this approach was employed by ByteDance’s TikTok, aiming to build a strong user base in the Global South. TikTok used the local content moderation approach in both Maghrebi Arabic and Kiswahili, where they tailored certain policies to align with regional and, at times, national standards and expectations. We observed that TikTok customized policies related to nudity, animal cruelty, and violence, adapting them from one region to another, or, in some cases, to groups of countries with similar cultural contexts. In the case of Maghrebi Arabic, we found that TikTok divides the Middle East and North Africa (MENA) region into three sub-regions: MENA 1 (Lebanon, Egypt, Palestine, Syria), MENA 2 (Sudan, Gulf, Yemen), and MENA 3 (Tunisia, Morocco, Algeria, Mauritania, and Libya), with MENA 1 and 2 recently being merged.

TikTok moderators we interviewed explained that they interpret policies differently depending on cultural nuances, such as how women's attire is viewed. For example, a video originating from a Gulf country showing a woman wearing short sleeves (MENA 2) would be removed, while the same video from Tunisia (MENA 3) would be allowed. One moderator, who worked for a vendor reviewing Arabic content for TikTok, shared that they were instructed to measure specific details, such as the number of centimeters of cleavage visible in a video. The decision to remove or keep content with visible cleavage was based on where the piece of content originated from and which rules apply to this country/region.

When we examined the Kiswahili case, we found that TikTok had different policies for each Kiswahili-speaking country. For instance, TikTok's drug policy allowed content about khat, a plant commonly chewed for its stimulant effects, to be posted by Kenyan content creators. However, such content was prohibited in Tanzania, and videos featuring khat from Tanzanian creators must be removed.

A few U.S.-based online forums and Indian social media companies employed a similar approach in Tamil. They altered policies and policy guidance to moderators to better fit the Tamil context. For example, one Indian social media company noted that speech in Tamil criticizing Periyar, an Indian philosopher commonly celebrated in South India, would not be permitted, although the same content in another Indian language would be. Some online forums enabled moderators to communicate with Tamil users through blog posts to explain how policies apply in Tamil and share feedback with policy teams in instances where moderation policies were being misapplied in Tamil contexts. In these contexts, companies hired native Tamil speakers and experts to moderate content and draft policies.

While this approach seems to acknowledge the cultural differences across regions, it can also severely limit freedom of expression, especially when users in the Global South attempt to challenge or push against local norms on social media. For instance, when moderation policies are tailored to fit with national norms, speech about LGBTQ+ issues, political issues, or women's rights may be suppressed should they not be permitted or seen as in line with the nation's cultural values, as is the case in many regions around the world (York, 2019). We often heard that Tamil users suspected their posts about specific elected officials or political parties were taken down or suppressed, particularly in instances when local governments would exercise authority under local legislation. Moderators too are given additional tasks when they must navigate and differentiate content policies across multiple countries.

The moderators we spoke with repeatedly mentioned how often they forgot which interpretations should be applied to which region. Many others also were apparently confused and could not remember the precise interpretations for different regions during our interviews.

Similar to the global approach, the assumption that moderators who speak a specific Global South language can effectively moderate all variations of that language has dominated the implementation of this approach in the Global South. There is a clear absence of recognition that each country and region has its own unique cultural and linguistic contexts in both the Global and Local approaches. In the Maghrebi Arabic case, moderators were assigned content from across the Arab world. For instance, a Tunisian moderator might be asked to review content from the Gulf region, a context and dialect they were less familiar with. In the Kiswahili case, outsourced vendors often hire only local moderators from the countries where they are based, primarily to avoid the complexities of visa and immigration regulations. As a result, the majority of the moderation is handled by Kenyans, with only a few Tanzanians involved. This has a negative impact on the quality of the moderation process, as moderators are often unfamiliar with the different regional contexts they are tasked with reviewing.

This was a significant problem in Quechua, which has a range of language varieties, some of which are very different from each other. Assuming that a few content moderators who speak Quechua can handle the range of varieties is short-sighted, yet some major social media companies take this approach. This is complicated by the reality that many users spell Quechua words in different ways, even within the same dialect, because the language is predominantly spoken, with few rigid rules in its written form. This lack of representation in the moderation teams may not only impact the accuracy of content moderation but also reflects a broader disregard for the linguistic and cultural nuances that are essential for accurate and sensitive content handling.

C. ALTERNATIVE PATHS: ONE LANGUAGE, MULTI-COUNTRY APPROACH

During our search for alternative approaches, we identified one approach that we found only in Kiswahili, employed by the local Tanzanian platform, JamiiForums. We refer to this as a "one language, multi-country" approach, as it goes beyond language-based moderation and focuses on moderating content according to each country's norms, context, and culture. JamiiForums primarily serves Swahili-speaking

users from diverse cultural backgrounds. The platform's leadership acknowledges that Kiswahili has multiple variations, and access to the local contexts where these language varieties exist is crucial to the firm's human rights mission. The forum relies less on automated content moderation and instead depends primarily on human moderators to minimize errors. However, this online forum has a limited user base of three million users, far fewer than platforms that implement global and local approaches.

Regardless of the approach a company decides to employ, outsourced vendors are key partners, making content moderation particularly challenging for low-resource languages. These vendors often prioritize commercial benefits over effectively serving people in these regions. In the following section, we explore how content moderation vendors exploit their workers and exacerbate the challenges of moderating content in these languages.

3. Manual and Automated Approaches to Enforcing Policy

Technology companies have long relied on poorly paid outsourced workers to produce both their hardware (smartphones and laptops) and software (data annotation) (Dvoskin et al., 2019). Over time, this labor force has grown to play a key role in the content moderation chain (Roberts, 2019). For example, Facebook alone hires more than 15,000 outsourced moderators to review about 3 million posts per day (Concentrix, 2025).

Companies that employ these workers are commonly referred to as “vendors,” “contractors,” “outsourcing agencies” (Dvoskin et al., 2019), or more formally as business process outsourcing companies (BPOs) (Carter, 2025). These terms emphasize the distance between workers and the tech companies they serve. Additionally, tech companies often make agreements with these vendors in secrecy, behind closed doors (Roberts, 2019). In many cases, users and the general public only learn about the labor behind content moderation when moderators speak out in the media or over social media, although that is rare due to strong contractual obligations and stigma.

In the early days of social media platforms, companies often employed in-house content moderators to review user-generated content. As one of our participants, who observed the early development of Facebook, explained, the platform initially relied on college students for content



▲ **Figure 7.** Mapping Business Processing Outsourcing companies who hire contract moderators responsible for content moderation in the four case studies we examined.

moderation. However, as social media platforms expanded globally, the practice of hiring in-house moderators became untenable and companies sought options that came at a lower financial cost, including by using automated moderation technologies to replace manual moderation and by recruiting moderators in regions where companies can pay less for the same service (Roberts, 2016). Consequently, tech companies shifted to contracting large global service providers that offer customer service, digital solutions, technical support, sales and marketing support, and other similar services that require large teams.

These service companies claim to be able to meet the demands of growing social media companies by offering large teams of employees. These companies have offices in many countries worldwide, particularly in the Global South, where they benefit from lower labor costs and weaker labor protections. These service companies promise training, health insurance, and decent paychecks for recent graduates or unemployed youth in contexts where employment opportunities are scarce. In many cases, recruits are not told directly what their role will be or what it entails — only that they’re applying for a remote role with little to no manual work.

BPO companies are often well-established and usually served many clients before they began providing content moderation. Tech companies sought to contract these BPOs to respond to the scale of

content moderation and the pressure from governments to take down content swiftly. The hundreds of millions of dollars in deals with BPOs remain secret and are not disclosed to the public. In our research, moderators shared that they were not informed about the nature of the job until they began their training. Moderators we interviewed mentioned that they had seen ads from these service companies hiring for roles in "customer service" or "data annotation," which did not clearly indicate that the positions were related to content moderation. Increasingly, as more content moderation becomes automated, these contract workers are tasked with roles powering automated moderation by labeling or annotating data to train keyword filters, large language models, and image matching tools ([Williams, 2022](#)).

When they applied, moderators, in the Kiswahili case, had to pass an English test and then pass an interview. In the Tamil case, one policy staffer responsible for overseeing moderators said that all moderators were required to understand English. This English requirement is needed to ensure that moderators are able to moderate English content from other countries and to read policies and interpretations written in English.

During job interviews, moderators are often only given limited information about the role. Many mistakenly believe they are being hired to become "content creators", while others initially assume they are being employed directly by the tech company. Our participants indicated that there were a few, if any, questions regarding the candidates' background or understanding of the region they would be moderating. Some of this is by design as many contractual obligations limit both BPO and moderators from disclosing which tech platforms they work for. The interviews often include a counselor asking superficial questions about the candidate's resilience, a question that many participants viewed as inadequate in assessing their mental health or readiness for the demands of the job. Moderators are not informed about the graphic content they will be reviewing or the amount of work assigned.

During the training, moderators start to learn more about the job. Details are given alongside the policies of the service which they must enforce. Moderators told us that these policies and guidance are far from the context they are familiar with. For example, former Quechua content moderators told us that the policies were drafted in Spanish and reflected a different set of cultural norms, often not relevant to Quechua content. As a result, they noted that there were several instances where harmful content in Quechua was not moderated because there were no Spanish equivalents to the terms being used and therefore no clear

policy violation to enforce. Moderators in these low-resource languages had little power to actually change the content policies in question, even if they were the resident experts in that language.

These policies are not publicly shared, and they are often too long for a moderator to learn in a few days. The training usually lasts for two weeks, followed by two tests: a dialect test and a policy test, both of which we were told are easy to pass. The training is usually based on examples from the Western world. Moderators told us that they were shown content from celebrity pages in the US and were taught how to review similar content. During the training, moderators are not shown the kind of graphic content they will encounter during their actual job. However, they are told about the positive impact of the job and that they are “the Internet Police,” “protecting their community” from harmful content. These kinds of messages pressured many of the moderators we spoke with to stick with the job, despite the trauma they experienced.

Moderators often realize the reality of this job only when they start working and are forced to balance the demand for efficiency with the impact of dealing with traumatic content, the lack of psychological support, and the challenge of addressing content in various dialects and contexts. Moderators we spoke with noted that they had to review from 700 to 1,000 tickets per day, or their performance metrics, and therefore income, would be negatively impacted. Moderators’ anxiety is not only related to traumatic content, but the speed at which they have to “action” these tickets. Moderators are timed and their success is measured in seconds ([Kgomo, 2025](#)). In some cases, well-being breaks are limited and deducted from the one-hour break allowed per day. If a moderator uses their one-hour break, no additional well-being break is permitted, regardless of how traumatizing the content they viewed.

The content moderation outsourcing business is expanding due to its high profitability. A recent report by the Everest Group’s Peak Matrix listed 27 companies as providers of Trust & Safety services ([Rickard et al., 2024](#)). In our research, four business outsourcing companies dominate the business of content moderation, particularly for low-resource languages (see Figure 7 and Table 3): Teleperformance, Sama, Majorel, and Concentrix. These four companies operate in four countries where we conducted our research: Kenya, Morocco, Tunisia, and Peru. In our Tamil case, we were unable to identify specific vendors. This does not indicate that no vendors are operating in the region, but rather that the platform representatives and moderators we consulted did not disclose such information.

Vendor	Languages Supported (from cases examined)	Operating Countries	Origin of the Company	Clients Served
Teleperformance	Maghrebi Arabic, Kiswahili, Quechua	Tunisia, Morocco, Kenya, and Peru	French	TikTok, Meta
Concentrix	Maghrebi Arabic	Tunisia	U.S.	TikTok
Sama	Kiswahili	Kenya	U.S.	Meta
Majorel (later acquired by Teleperformance)	Maghrebi Arabic, Kiswahili, and Quechua	Tunisia, Morocco, Kenya, and Peru	Luxembourg	TikTok, Meta

▲ **Table 3.** Vendors offering content moderation services for low-resource languages (from the cases we examined). Note: Majorel was later acquired by Teleperformance.

Platforms also employ automated moderation technologies to enforce policies and moderate content at scale. These include keyword filters, machine learning-trained classifiers, and hash-matching tools to remove images containing nudity (Shenkman et al. 2021). Their use is increasingly necessary because of the scale at which companies operate and the proliferation of laws that require platforms to take down illegal content with short notice (Gorwa et al., 2020). However, these systems are known to be — and were perceived by interviewees to be — error-prone across cases.

We spoke with platform representatives, moderators, and users to understand their perception of moderation broadly as well as automated moderation specifically. We also hosted roundtables with NLP experts to understand the unique challenges and opportunities for building automated technologies in the languages we studied. Online platforms often noted that it was difficult and expensive to build automated moderation tools in low resource languages because of the dearth of training data and research.

NLP researchers and language technology experts, mostly from academia, reiterated that low availability of training data representative of how users speak online in these languages makes it challenging to moderate content effectively and accurately. Users speak bilingually at times, mixing their native tongues and other adjacent or dominant languages. Sometimes they code-switch or code-mix. Other times they use algospeak, developing linguistic tricks to evade moderation, such as using “unalive” to talk about suicide or the watermelon emoji to refer

to Palestine ([Lorenz, 2022](#); [Elswah, 2024a](#)). These types of terms often are not represented in training datasets. In many cases, unique in-group slurs are also ignored despite widespread research documenting them.

We found that models do not perform well with Maghrebi Arabic dialects compared to other popular dialects in the region. Additionally, there are limited datasets on Arabizi (using Latin letters to write Arabic words) and code switching (using Arabic and other languages in one sentence) to train the classifiers. Moreover, we found that the limited hiring of annotators who speak the native dialects of the Maghrebi Arabic region has led to inconsistencies, bias, and inaccuracies in labeled datasets, which ultimately impact the quality of the classifiers.

Similarly, we found that the evolution of the Kiswahili language, along with the growing popularity of Sheng (a Kiswahili variation used by young adults) and code mixing, have also led to complexities similar to those we found in the Maghrebi Arabic case. These factors contributed to further complexities in data collection and model generalization. Sheng, for example, is rapidly evolving with new words and phrases emerging frequently, presenting unique challenges for NLP models compared to more stable dialects. As in the Maghrebi Arabic case, the lack of resources and funding have also led to critical issues with data access and annotation. NLP researchers and language technology experts in the Kiswahili contexts have repeatedly relied on students or AI tools to annotate the data.

We found that companies have moved to automated systems including large language models for both Quechua and Spanish, even though there are vast differences in how much each of those languages is resourced. One representative from a social media company reported that in some cases they actually machine-translated content from Quechua to Spanish for review. Many of the linguists, NLP researchers, and LLM developers we spoke to were surprised that social media companies would rely on these kinds of automated systems for content in Quechua, believing that the technologies were not yet ready for this use case. They argued that a major problem is the lack of investment in research and development of training data for Quechua, particularly in relation to Spanish. Training data is particularly relevant given the numerous varieties that exist in this language. Among many of the researchers we spoke to, this trend reflected the ongoing privileging of Spanish over Quechua, a problem that has existed since colonial times.

In the Tamil case, too, roundtable participants argued that more training data was needed to better represent the regional variations of the language, the breadth of the vocabulary, and the unique Tamlish and Tanglish variations which include users writing Tamil

in Romanized script and using both Tamil and English in the same sentence. Tamil is also a diglossic language, which means there is a formal written Tamil and a spoken one, the latter of which is often transliterated online to enable more colloquial discussions. Often this distinction is not apparent in training corpus, with a lot of it coming from more formal sources (i.e., government documents or classical literature).

As previously mentioned, users have developed novel and sophisticated techniques to circumvent moderation, and this often affects the efficacy of automated content moderation systems. Algospeak and other forms of computer-mediated communication are often not represented in training datasets and are therefore incorrectly moderated.

More data and evaluation resources are also needed across the four languages to robustly test these systems. Platforms rely upon machine-translated benchmarks that are not adequately scoped to multilingual and multicultural contexts and miss out on shortcomings of these systems. Moderating things like hate speech and gender-based violence often requires deep cultural knowledge and an understanding of local contexts, yet concepts outside the Global North are often not represented in bias evaluation datasets.

Lack of prioritization and Anglocentric assumptions also result in poor capabilities of automated systems and inconsistent outcomes in content moderation. Users interviewed across the four case studies felt that platforms did not care about them and that they experienced moderation differently when they wrote in English or Spanish compared to their native language. This is reflected in an interview we had with one user who sent requests for an automated captioning service to improve its Tamil-language capabilities and was told that this was not a priority for the U.S.-based company. Similarly, another interviewee, a Quechua content creator, reported that captioning was not available for the music videos they uploaded; because they created content in part to share their culture and language with others, this was an ongoing problem.

Further, despite civil society and academic researchers' work improving low resource lexicons and automated tools, few companies seem to adopt these resources, allowing caste-based and gender-based slurs to proliferate widely on social media. Some users believe that Western platforms simply have more tolerance for hate speech and harassment in Global South contexts, meaning that investment and enforcement thresholds are higher than in other contexts.

4. Users experienced moderation differently and inconsistently

Through interviews, we learned that many users experienced what they perceived as inconsistent moderation in which other users' content they believed was violating was left up, yet content they posted which they believed did not violate terms of service was taken down. In one case, an interviewed user noted that harassment targeting them was left up despite their repeated reporting of the post. Yet, when they posted screenshots of the messages and comments they received, they were taken down.

Marginalized users perceived more inconsistent moderation than other users. Some users believe they faced targeted harassment because they posted in a non-English language online. For example, interviewees said that they faced targeted harassment because they spoke Quechua, which reflected the racism and discrimination that Quechua-speaking communities have faced for hundreds of years. They also described it as part of the cost of engaging on social media.

In the Tamil case, individuals who identify as LGBTQ+ or belonging to a caste that was traditionally oppressed reported facing more targeted harassment online. When they tried to report these posts, they would not receive a response from the company. They also noted that when well-known slurs were used to target them, they were not flagged or taken down even when they apparently violated a platform's terms of service, despite the swift removal of words with similar connotations shared in different languages b. To some users this indicated that harassment of these communities was tolerated or neglected by platforms.

5. Tactics of Resistance & Fighting Back

Data from the four case studies demonstrates that users employ sophisticated tactics and exemplify resistance to what they see as poor moderation efforts in order to enjoy access to a fair and equal digital experience.

Throughout our interviews, roundtable sessions, focus group sessions, and informal meetings with users, digital rights advocates and content creators, we observed a consistent theme: the use of innovative tactics to evade what they perceived as burdensome content moderation. In general, we found three main tactics to evade moderation: algospeak, mass reporting, and civil society interventions (See Table 4). Using these

Language	Tactic	Tactic Type	Description
Maghrebi Arabic	Mass Reporting	Mobilizing and Organizing Other Users	Working with large groups of users to report content or accounts that the social media companies have failed to remove using standard reporting channels in an effort to get those companies to delete accounts and pages.
	Algospeak	Randomizing Content	Uploading random content alongside material users believe will be restricted.
		Spelling Substitution	Substituting letters with numbers or employing other unconventional spelling techniques like dotless Arabic.
		Symbol/Emoji Substitution	Using symbols or emojis as substitutions for other concepts; one key example is the use of the watermelon emoji as a reference to the Palestinian flag.
		Avoidance of Words	Avoiding altogether certain words likely to trigger moderation.
	Civil Society Intervention	Escalating the Issue	Civil society organizations in the region have taken on the responsibility of acting as the “middle man,” connecting concerned users with tech companies, especially in relation to takedowns and restriction of harmful content.
Kiswahili	Mass Reporting	Mobilizing and Organizing other users	Relying on social capital mobilizes other users to try to get platforms to address content they were unable to remove through official reporting mechanisms.
	Civil Society Intervention	Escalating the Issue	Relying on third-party escalation and civil society groups to report harmful content in some cases.
Tamil	Algospeak	Latin Letter Integration	Code-mixed and transliterated Tamil using Latin script for ease and convenience.
			Cropping images, blurring symbols pertaining to the Sri Lankan civil war.
		Spelling Substitution	Using asterisks
Quechua	Post in Quechua	Alternating the Language	Posting in Quechua to avoid automated moderation.

▲ Table 4. Tactics of resistance.

methods, users attempt to regain control over their experience with opaque and inconsistent moderation processes.

In the Maghrebi Arabic case, users employed many different tactics in order to counter what they saw as poor, inconsistent moderation. Some of this was due to the urgent need and desire to share content related to Israel's invasion of Gaza despite increasing takedowns and content restrictions during the period in which we collected data (April to September 2024). This increase in takedowns has also been identified by organizations such as Human Rights Watch, 7amleh, and others. Users employed several tactics, including using "algotalk," by randomizing the order of their speech, changing the spelling of words or terms, and using symbols or emojis in order to bypass content filters. Many interviewees suggested they used the watermelon emoji when they referred to Palestine to evade automated content moderation.

Similarly, in the Tamil case study, many users strategically cropped or blurred images or symbols associated with political entities or taboo themes in order to bypass content filters. In many cases, Tamil users felt their speech was taken down if they mentioned political entities or elected officials. To circumvent this, many users developed tactics such as adding asterisks or spaces in words, believing that these filters were created to detect specific keywords but not sophisticated enough to understand the Tamil script. For example, users would elongate Tamil words by adding spaces between characters or flip the order of Tamil characters, rendering them nonsensical to content filters. Nonetheless, Tamil speaking users still reported enormous levels of inconsistent moderation and repeated account suspensions, particularly when they were speaking about political issues or sharing news from particular news outlets.

Users often reported facing more harassment when speaking in Quechua than in Spanish, although they often code-mixed. Several Quechua interviewees noted different standards of moderation for Spanish and Quechua. In many cases, Quechua content would go unmoderated because of what they suspected was a lack of understanding of the language by social media companies. Given this experience, some interviewees said that one way to express speech likely to be moderated by the platforms in Spanish was simply to post in Quechua.

The use of Quechua online is itself a form of resistance given the long history of suppression of the language and its people. In fact, many Quechua social media users generate content with a focus on sharing their language and culture with other Quechua communities and the

world ([Mendoza-Mori & Sanchez, 2023](#)). Many of the interviewees we spoke to mentioned that posting in Quechua helps to promote a Quechua identity in different spaces ([Thakur, 2025](#)). This is part of a larger trend of indigenous peoples using various media as a means of self-representation, cultural agency, and resistance ([Pérez, 2020](#)).

Additionally, users in the Global South relied on mass reporting techniques to get platforms to remove content or pages they deemed harmful or in violation of company policies, particularly when they were unable to do so through traditional reporting channels. This tactic, which first emerged during the Arab Spring in 2011 ([Elsawah, 2023](#)), continues to be used, whether to silence others or to push back against perceived violations. The tactic involves mobilizing and organizing a large number of users to simultaneously report a specific page or piece of content until it is removed. We found this tactic to be prevalent in both Maghrebi Arabic and Kiswahili, though less so in Tamil and Quechua.

Lastly, users turned to civil society organizations to escalate issues related to content restrictions and removals. Civil society in the Global South plays a vital role as a bridge between users and tech companies. Many civil society organizations act as trusted partners to social media companies or have strong ties with employees and tech company representatives, allowing them to directly escalate users' issues. The continued efficacy of this path is unclear due to layoffs of critical Trust & Safety teams within tech companies ([McCorvey, 2023](#)). Some civil society organizations offer dedicated helplines to be more reachable to users. Despite this, digital rights advocates we interviewed expressed frustration with the slow and sometimes non-responsive nature of companies when addressing the issues. In some cases, these channels have failed to achieve their objectives, with certain trusted partner programs lacking the responsiveness needed ([Internews, 2023](#)).

We found this tactic adopted in the Maghreb region, where, compared to other parts of the Arab world, a strong network of civil society organizations facilitates these escalations. Many participants we interviewed for the Kiswahili case also referred to the importance of these escalations. However, this type of civil society presence is not available in many other countries in the Global South.



Recommendations



1. Disclose more to users about when, why, and how moderation occurs, and use civil society and user feedback to influence policy development and enforcement.

Content moderation always includes tradeoffs. However, for users in the Global North, erroneous moderation can often be clearly mapped onto a policy or terms of service. For users in the Global South, visibility into why a moderation decision was made and the policies that enabled or legitimized it is often more difficult. This is because platform policies often are not translated, are translated infrequently, or have critical gaps pertaining to the local context. As such, individual users face barriers in understanding why a piece of content is taken down and in remedying decisions when policies don't adequately capture the unique local context.

Enhancing transparency and opportunities for remedy so that users around the world are able to understand a company's policies will breed not only trust but also greater good faith participation on platforms. In the absence of transparency, and consequently belief and trust with moderation processes, users feel it is necessary to evade moderation practices to speak freely, further complicating effective content moderation. As part of this transparency, platforms should make clear whether a piece of content was moderated using a human reviewer or an automated moderation tool, whether that automated tool was trained on the language of the post, the policy the user violated, and how the user can appeal a decision made by the platform.

2. Ensure the adoption of linguistic and cultural expertise in all parts of the development and implementation of content moderation policies and processes.

Regional, linguistic, and cultural diversity is necessary at all stages of policy development, enforcement, and tool development when it pertains to content moderation. The nature of linguistic and cultural expertise will differ based on where in the Trust & Safety value chain it occurs. But the necessity of linguistic and cultural expertise has been laid bare across the four case studies, particularly in instances highlighting

how ignoring linguistic and cultural nuances can impede user safety, rights, and moderation efficacy.

Accommodating linguistic and cultural expertise can look like engaging with communities and conducting user research on how policy enforcement differs from language to language. Many users across the four case studies mentioned that social media companies had never engaged their community before, particularly in the Quechua case study, despite their efforts to engage with company regional teams or moderation efforts. Engaging with these communities not only tangibly improves moderation by making sure policies and policy enforcement processes adequately considers unique linguistic considerations. It also helps build trust within these communities so that moderation is effective and not circumvented inappropriately.

Moreover, ensuring adequate diversity and expertise at earlier stages (such as when developing or stress-testing policies and developing tools to enforce them) is critical to ensuring moderation works effectively. Additionally, ensuring that automated systems are developed in concert with NLP research consortia is critical to making sure automated systems can adequately parse non-English text. NLP researchers, including the ones that we have spoken with, have built datasets, techniques, and evaluation methods to ensure moderation practices work equitably and efficiently across languages. Their tools should be used by companies to ensure moderation practices work robustly.

3. Better working conditions for content moderators.

Moderators currently endure challenging working conditions that have left many of them traumatized and emotionally distressed. This distress significantly impacts their daily content moderation decisions. These decisions not only affect moderators personally but also play a crucial role in the content moderation process. Accuracy is essential, not only to ensure fair treatment of all content but also because these decisions serve as training data for automated moderation models.

Content moderation vendors should offer better psychological support, improved compensation, and regular wellness breaks for their moderators. Performance evaluations should emphasize decision accuracy over the sheer volume of tickets processed. Additionally, there is a pressing need for more extensive training before moderators begin their work. Current practices, which typically offer only two weeks

of training, are inadequate according to feedback from moderators. Moreover, moderators should have a direct line to communicate with tech company policy teams. Feeling marginalized and unheard undermines effectiveness and morale within the moderation ecosystem. Lastly, moderators deserve the ability to openly discuss their work experiences and share their challenges with colleagues and loved ones. The current culture of secrecy in content moderation is harmful to both moderators and the industry as a whole.

4. Reshaping current content moderation approaches.

Current content moderation approaches often fall short because they either apply global, one-size fits all policies or overly localized approaches that might limit users. Tech companies should adopt a new model of content moderation that respects the cultural nuances and linguistic diversity of each region. This includes creating an opportunity for users to challenge local norms that violate principles of free expression or other human rights. Companies must develop a participatory-based approach that is shaped by insights from users in Global south regions. This also includes hiring a diverse group of moderators that can cover and review different dialects from a single region.

Some companies have succeeded by moving beyond the global vs. local debates. These companies recruited a diverse group of moderators and policy teams, involved users, and offered more effective communication channels with users. These small-scale attempts offered a more sensitive approach than just adherence to local norms and customs. Such attempts should become the norm and institutionalized across online services to ensure adequate moderation in the Global South.

5. Develop AI Models Tailored to Regional Linguistic Variations.

As the scale of content posted on social media platforms continues to increase, reliance on automated moderation techniques also continues to increase. Improvements to automated capabilities are essential to ensure speech is moderated consistently, as is ensuring human oversight of automated moderation. As the four case studies show, there are few effective automated tools available in many low resource languages, and the ones that are currently being employed are often error prone. NLP

experts who spoke with us for these case studies highlighted critical steps that are required to improve automated moderation systems in these languages.

First, companies should try to source or develop natively created datasets to train automated tools. Some of these datasets are already available and accessing them is a matter of asking NLP experts building these resources around the world and developing an internal pipeline to identify, vet, and implement these independently created resources. Second, companies should ensure that evaluation metrics used to test existing automated moderation tools are developed with NLP experts familiar with the language in question and the regional variations of the language. As identified across the four case studies, the way low resource languages are spoken differ from region to region. Ensuring that these systems are tested across language and region will ensure systems work as planned. Finally, automated systems should be repeatedly tested using natively created evaluation sets to ensure that as language evolves these systems are working as planned.



Appendix



Methods & Data Collection

To study the state of content moderation of low-resource languages in the Global South, we used a mixed-methods approach, combining qualitative and quantitative methods. Our approach in studying this issue was to adapt to any methodological complexity or challenge. In the Maghrebi Arabic case, our first case, we conducted five focus group sessions with 25 content creators and users in the region. Through these focus group sessions, we were able to get to know users’ day-to-day uses of online services, the obstacles they face, their knowledge of the community standards, and how they evade algorithms. Despite that, we decided to discontinue the focus groups for the case studies replacing them with in-depth interviews with content creators and users from the region. After adopting this change in method, we realized how necessary it was to hear detailed accounts of each users’ perspective about content moderation.

Across the four cases, we conducted semi-structured interviews with 69 experts to assess the information environment, learn more about the state and perception of content moderation, and understand firsthand how content moderation works and where it falls short. We spoke with Trust & Safety and policy professionals at Western as well as regional companies, including moderators tasked with reviewing speech from different Global South contexts, digital rights experts, academics, and content creators.

Most interviews were conducted online, except for one in-person interview during a field visit in Bangalore, India. Interviews were conducted between March 2024 and April 2025. The majority of the interviews were held in English, however where necessary interviews were carried out in Spanish, Arabic, and Hindi. Field notes were taken during the interviews and recordings of the interviews were later transcribed and analyzed to find common themes.

In addition to interviews, we also conducted an online survey to understand the experience of 562 frequent social media users in the Global South. We relied on our regional partners, Digital Citizenship, Paradigm Initiative, Center for Internet and Society, and Hiperderecho to disseminate the survey locally and in surrounding countries.. Survey data was collected from April 2024 to March 2025. In each region, we translated the survey into the native languages of the target countries to ensure broader representation of regional social media users.

The survey asked questions about users' trust in social media platforms, their experiences facing moderation, and their perceptions of moderation. We used the Alchemer survey platform to distribute the online survey and offered a modest honorarium of US\$10 for participation. As described above, the survey was made available in English, Arabic, French, Kiswahili, Tamil, Spanish, and in three varieties of Quechua — (Central Quechua (Wanka, Ancash, Pataz), Southern Quechua (Chanka, Collao), and Quichua/Kichwa (Ecuador, Amazon)). Our respondents represent the following countries: Tunisia, Morocco, Algeria, Tanzania, Kenya, Peru, Bolivia, Ecuador, India and Sri Lanka.

Finally, we held four virtual roundtables with 39 computer science experts, Natural Language Processing researchers, and language technology experts who were familiar with building, evaluating, and annotating datasets for machine learning technologies. These discussions sought to highlight the challenges various types of specialized experts face in data collection, annotation, evaluation, and to circulate solutions identified within these communities to address these specific challenges.



References

- Alimardani, M., & Elswah, M. (2021, August 5). Digital Orientalism: #SaveSheikhJarrah and Arabic Content Moderation. *Project on Middle East Political Science*. <https://pomeps.org/digital-orientalism-savesheikhjarrah-and-arabic-content-moderation> [perma.cc/6ZVE-DA8S]
- Arshat, A., & Etcovitch, D. (2018). The Human Cost of Online Content Moderation. *Harvard Journal of Law & Technology*. <https://jolt.law.harvard.edu/digest/the-human-cost-of-online-content-moderation> [perma.cc/ZJ7V-MR89]
- Badouard, R., & Bellon, A. (2025). Introduction to the special issue on content moderation on digital platforms. *Internet Policy Review*. <https://doi.org/10.14763/2025.1.2005> [perma.cc/D7JY-2ZL9]
- Barnes, M. R. (2022). Online Extremism, AI, and (Human) Content Moderation. *Feminist Philosophy Quarterly*, 8(3/4), Article 3/4. <https://doi.org/10.5206/fpq/2022.3/4.14295> [perma.cc/Q26N-2G2R]
- Bhatia, A., & Elswah, M. (2025, May 14). Moderating Tamil Content on Social Media. *Center for Democracy and Technology*. <https://cdt.org/insights/moderating-tamil-content-on-social-media/> [perma.cc/G9UF-96ZC]
- Carter, R. (2025, January 10). The Top Business Process Outsourcing Companies for 2025. *CX Today*. <https://www.cxtoday.com/workforce-engagement-management/the-top-business-process-outsourcing-companies/> [perma.cc/37LW-279R]
- Cieri, C., Maxwell, M., Strassel, S., & Tracey, J. (2016). Selection Criteria for Low Resource Language Programs. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 4543–4549). European Language Resources Association (ELRA). <https://aclanthology.org/L16-1720/> [perma.cc/A3DS-5RNY]
- Crystal, C. (2023). *Facebook, Telegram, and the Ongoing Struggle Against Online Hate Speech*. Carnegie Endowment for International Peace. <https://carnegieendowment.org/research/2023/09/facebook-telegram-and-the-ongoing-struggle-against-online-hate-speech?lang=en> [perma.cc/XV2Q-W26L]
- Duarte, N., Llansó, E., & Loup, A. C. (2017). *Mixed Messages? The Limits of Automated Social Media Content Analysis*. Center for Democracy & Technology. <https://cdt.org/wp-content/uploads/2017/11/2017-11-13-Mixed-Messages-Paper.pdf> [perma.cc/9QJ4-BTTZ]
- Dvoskin, B. (2023). What the United Nations Guiding Principles for Business and Human Rights (Don't) Say About Content Moderation. *Lawfare*. <https://www.lawfaremedia.org/article/what-united-nations-guiding-principles-business-and-human-rights-dont-say-about-content-moderation> [perma.cc/685K-J5EM]
- Dvoskin, E., Whalen, J., & Cabato, R. (2019, July 25). Content moderators at YouTube, Facebook and Twitter see the worst of the web—And suffer silently. *The Washington Post*. <https://www.washingtonpost.com/technology/2019/07/25/social-media-companies-are-outsourcing-their-dirty-work-philippines-generation-workers-is-paying-price/> [perma.cc/H9PQ-FHF5]
- Dzahene-Quarshie, J. (2009). Globalization of an African Language: *Truth or Fiction?* <https://ugspace.ug.edu.gh/items/ccca20e1-b8d7-4ff0-adfb-0ddc1b13fc2e> [perma.cc/D3XJ-F965]

- Egbunike, N. (2020, October 28). How Facebook derailed Nigeria's #EndSARS movement. *Global Voices Advox*. <https://advox.globalvoices.org/2020/10/28/how-facebook-derailed-nigerias-endsars-movement/> [perma.cc/9ANN-4J4A]
- Elsawah, M. (2023). Online tactical innovation and stagnation: *Insights from the aftermath of the Arab spring in Syria and Tunisia* [Http://purl.org/dc/dcmitype/Text, University of Oxford]. <https://ora.ox.ac.uk/objects/uuid:bdf2c6f7-2f2c-4488-9480-14852369e985> [perma.cc/UE6N-7HJW]
- Elsawah, M. (2024a). *Moderating Maghrebi Arabic Content on Social Media*. Center for Democracy & Technology. <https://cdt.org/insights/moderating-maghrebi-arabic-content-on-social-media/> [perma.cc/M2LM-DK5J]
- Elsawah, M. (2024b). *Moderating Kiswahili Content on Social Media*. Center for Democracy & Technology. <https://cdt.org/insights/moderating-kiswahili-content-on-social-media/> [perma.cc/9AW6-A6HM]
- Elsawah, M. (2024c, October 29). *How to Moderate Cleavage on Social Media?* | *TechPolicy.Press*. Tech Policy Press. <https://techpolicy.press/how-to-moderate-cleavage-on-social-media> [perma.cc/Z2CE-7FRB]
- Etter, L. (2017, December 7). Rodrigo Duterte Turned Facebook Into a Weapon, With a Little Help From Facebook. *Bloomberg.Com*. <https://www.bloomberg.com/news/features/2017-12-07/how-rodrigo-duterte-turned-facebook-into-a-weapon-with-a-little-help-from-facebook> [perma.cc/F5A3-29K2]
- France 24. (2023, August 3). *Zimbabwe election disinformation spreads on WhatsApp*. France 24. <https://www.france24.com/en/live-news/20230803-zimbabwe-election-disinformation-spreads-on-whatsapp-1> [perma.cc/7UGT-5Q6P]
- Gillespie, T. (2018). *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media*. Yale University Press. [perma.cc/9A4Z-D9AC]
- Gillespie, T. (2020). Content moderation, AI, and the question of scale. *Big Data & Society*, 7(2), 2053951720943234. <https://doi.org/10.1177/2053951720943234> [perma.cc/R33X-62GQ]
- Gorwa, R., Binns, R., & Katzenbach, C. (2020). Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1), 2053951719897945. <https://doi.org/10.1177/2053951719897945> [perma.cc/K3NH-H84C]
- Grimmelmann, J. (2015). The Virtues of Moderation. *Cornell Law Faculty Publications*. <https://scholarship.law.cornell.edu/facpub/1486> [perma.cc/SX8P-7A2V]
- Harrat, S., Meftouh, K., & Smaïli, K. (2018). Maghrebi Arabic dialect processing: An overview. *Journal of International Science and General Applications*, 1. <https://hal.science/hal-01873779> [perma.cc/9LAH-2VY7]
- Internews. (2023). Safety at Stake: How to Save Meta's Trusted Partner Program. *Information Saves Lives* | Internews. <https://internews.org/resource/safety-at-stake-how-to-save-metas-trusted-partner-program/> [perma.cc/8UN9-CN74]
- Jalli, N. (2023, March 23). *How TikTok became a breeding ground for hate speech in the latest Malaysia general election*. *The Conversation*. <http://theconversation.com/how-tiktok-became-a-breeding-ground-for-hate-speech-in-the-latest-malaysia-general-election-200542> [perma.cc/RV5D-M79F]


- 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/5JNN-H67L]
- Kamara, S., Thakur, D., Post, G., Knodel, M., Llansó, E., Nojeim, G., Vogus, C., & Llanso, E. (2021, August 12). Outside Looking In: Approaches to Content Moderation in End-to-End Encrypted Systems. *Center for Democracy and Technology*. <https://cdt.org/insights/outside-looking-in-approaches-to-content-moderation-in-end-to-end-encrypted-systems/> [perma.cc/JTB5-LM5B]
- Kgomo, S. (2025, February 12). I was a content moderator for Facebook. I saw the real cost of outsourcing digital labour. *The Guardian*. <https://www.theguardian.com/commentisfree/2025/feb/12/moderator-facebook-real-cost-outsourcing-digital-labour> [perma.cc/F6VJ-NXBJ]
- Lorenz, T. (2022, April 8). Internet ‘algospeak’ is changing our language in real time, from ‘nip nops’ to ‘le dollar bean.’ *The Washington Post*. <https://www.washingtonpost.com/technology/2022/04/08/algospeak-tiktok-le-dollar-bean/> [perma.cc/U7J6-3LQN]
- Mahmud, T., Ptaszynski, M., Eronen, J., & Masui, F. (2023). Cyberbullying detection for low-resource languages and dialects: Review of the state of the art. *Information Processing & Management*, 60(5), 103454. <https://doi.org/10.1016/j.ipm.2023.103454> [perma.cc/RG9K-DXQD]
- McCorvey, J. (2023, February 10). Tech layoffs shrink ‘trust and safety’ teams, raising fears of backsliding efforts to curb online abuse. *NBC News*. <https://www.nbcnews.com/tech/tech-news/tech-layoffs-hit-trust-safety-teams-raising-fears-backsliding-efforts-rcna69111> [perma.cc/DNE3-5FC7]
- Mendoza-Mori, A., & Sanchez, M. A. B. (2023, April 25). *Quechuactivism in Social Media: Digital Content and Indigenous Language Awareness* | *ReVista*. <https://revista.drclas.harvard.edu/quechuactivism-in-social-media-digital-content-and-indigenous-language-awareness/> [perma.cc/H39L-48VY]
- 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/RLR3-9HGA]
- Mwaura, J. (2023). Silicon Savannah or Digitising Marginalisation?: A Reflection of Kenya’s Government Digitisation Policies, Strategies, and Projects. In *Communication Rights in Africa*. Routledge. [perma.cc/G3AF-AD4K]
- Nandakumar, T., & Amarasingam, A. (2021, May 14). *Social Media Platforms are Silencing Social Movements* | *TechPolicy.Press*. [Tech Policy Press. https://techpolicy.press/social-media-platforms-are-silencing-social-movements](https://techpolicy.press/social-media-platforms-are-silencing-social-movements) [perma.cc/7LQX-DB29]
- Nicholas, G., & Bhatia, A. (2023). 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/8XQ5-9YE4]
- Pérez, V. L. (2020). El translenguaje digital, estrategia discursiva ecológica de jóvenes bilingües quechua–Castellano en Facebook y Whatsapp. *Ecolinguística: Revista Brasileira de Ecologia e Linguagem (ECO-REBEL)*, 6(1), 83–103. <https://periodicos.unb.br/index.php/erbel/article/view/29898> [perma.cc/WQK8-7BEQ]


- Concentrix (2025). *The Ultimate Guide to Content Moderation*. Concentrix. <https://www.concentrix.com/insights/blog/the-ultimate-guide-to-content-moderation/> [perma.cc/C2V8-DUQE]
- Renaldi, A. (2024, July 23). *In Indonesia, social media is a “bunting ground” for religious minorities*. Rest of World. <https://restofworld.org/2024/indonesia-muslim-influencers/> [perma.cc/3W7P-BS92]
- Rickard, D., Dasgupta, A., Sengupta, A., Jain, S., Khosla, D., Pathela, T., & Shaw, S. (2024). *Everest Group Reports—View*. https://www2.everestgrp.com/reportaction/EGR-2024-67-R-6387/Marketing?_gl=1%2Acvyqt%2A_gcl_au%2AMTAxMDA1NDQ0My4xNzQ2MzU0MTk3%2A_ga%2AMTUzODk3MDIyOS4xNzQ2MzU0MTk3%2A_ga_DPVMZLDPS2%2AczE3NDYzNTQxOTckbzEkZzAkdDE3NDYzNTQxOTckajYwJGwwJGgw [perma.cc/AFB5-98SK]
- Roberts, S. (2016). *Commercial Content Moderation: Digital Laborers’ Dirty Work*. Media Studies Publications. <https://ir.lib.uwo.ca/commpub/12> [perma.cc/WJ5P-BUQW]
- Roberts, S. (2019). *Behind the Screen*. <https://yalebooks.yale.edu/book/9780300261479/behind-the-screen/> [perma.cc/M3A2-45XV]
- Robinson, N. R., Ogayo, P., Mortensen, D. R., & Neubig, G. (2023). *ChatGPT MT: Competitive for High- (but not Low-) Resource Languages* (arXiv:2309.07423). arXiv. <https://doi.org/10.48550/arXiv.2309.07423> [perma.cc/5FLL-4P52]
- Sarveswaran, K. (2024). *Tamil Language Computing: The Present and the Future* (arXiv:2407.08618). arXiv. <https://doi.org/10.48550/arXiv.2407.08618> [perma.cc/M7GH-8VSF]
- Sengupta, A., Ripstra, A., & Vrana, A. (2020). *State of the Internet’s Languages report*. <https://internetlanguages.org/en/summary/> [perma.cc/2RSR-NQQN]
- Shahid, F., Elswah, M., & Vashistha, A. (2025). *Think Outside the Data: Colonial Biases and Systemic Issues in Automated Moderation Pipelines for Low-Resource Languages* (arXiv:2501.13836). arXiv. <https://doi.org/10.48550/arXiv.2501.13836> [perma.cc/M8RR-NFN5]
- Shenkman, C., Thakur, D., & Llansó, E. (2021). *Do You See What I See? Capabilities and Limits of Automated Multimedia Content Analysis*. Center For Democracy And Technology. [perma.cc/GRX8-X77Q]
- 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/42YF-QK74]
- Thakur, D. (2025). *Moderating Quechua Content on Social Media*. Center For Democracy And Technology. <https://cdt.org/insights/moderating-quechua-content-on-social-media/> [https://perma.cc/ENS6-DZ6U]
- Wahome, M. N. (2023). Introduction: ‘Shooting for the Moon.’ In M. N. Wahome (Ed.), *Fabricating Silicon Savannah: The Making Of A Digital Entrepreneurship Arena Of Development* (pp. 1–11). Springer International Publishing. https://doi.org/10.1007/978-3-031-34490-9_1 [perma.cc/XWH9-K8VN]

- Williams, A. (2022). *The Exploited Labor Behind Artificial Intelligence*. <https://www.noemamag.com/the-exploited-labor-behind-artificial-intelligence> [<https://perma.cc/SFX6-8VQT>]
- Wolk, R. M. (2004). The effects of English language dominance of the Internet and the digital divide. *2004 International Symposium on Technology and Society*, 174–178. <https://doi.org/10.1109/ISTAS.2004.1314348> [perma.cc/8VWC-HMFC]
- York, J. C. (n.d.). *Silicon Values*. Verso. Retrieved June 14, 2025, from <https://www.versobooks.com/en-gb/products/882-silicon-values> [perma.cc/JV3M-4UZ6]

 cdt.org

 cdt.org/contact

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

 202-637-9800

 @CenDemTech

