MediaWell

News Item

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

By Mona Elswah, Aliya Bhatia, and Dhanaraj Thakur | Center for Democracy & Technology (CDT)
September 9, 2025

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 (Elswah, 2024a), Kiswahili (Elswah, 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 over 560 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 is known 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.

Read the full report.