Proceedings of the Eighth AAAI/ACM Conference on AI, Ethics, and Society (AIES 2025)

Think Outside the Data: Colonial Biases and Systemic Issues in Automated Moderation Pipelines for Low-Resource Languages

Farhana Shahid[1], **Mona Elswah**[2, 3], **Aditya Vashistha**[1]
[1]Cornell University
[2]Center for Democracy and Technology
[3]University of Exeter
fs468@cornell.edu, m.a.i.elswah@exeter.ac.uk, adityav@cornell.edu

Abstract

Most social media users come from the Global South, where harmful content usually appears in local languages. Yet, AI-driven moderation systems struggle with low-resource languages spoken in these regions. Through semi-structured interviews with 22 AI experts working on harmful content detection in four low-resource languages: Tamil (South Asia), Swahili (East Africa), Maghrebi Arabic (North Africa), and Quechua (South America)–we examine systemic issues in building automated moderation tools for these languages. Our findings reveal that beyond data scarcity, socio-political factors such as tech companies' monopoly on user data and lack of investment in moderation for low-profit Global South markets exacerbate historic inequities. Even if more data were available, the English-centric and data-intensive design of language models and preprocessing techniques overlooks the need to design for morphologically complex, linguistically diverse, and code-mixed languages. We argue these limitations are not just technical gaps caused by "data scarcity" but reflect structural inequities, rooted in colonial suppression of non-Western languages. We discuss multi-stakeholder approaches to strengthen local research capacity, democratize data access, and support language-aware solutions to improve automated moderation for low-resource languages.

1 Introduction

The largest and fastest-growing social media user base comes from the Global South, where billions generate content in their local languages. This growth has fueled misinformation, hate speech, and incitement to violence in non-English languages, contributing to serious human rights violations (Samuels 2020; Milmo 2021; Yibeltal and Muia 2023). However, the moderation systems deployed by tech companies prioritize English-speaking users in the West (Legon and Alsalman 2020; Popli 2021), leaving harmful content in Global South languages largely unchecked, increasing social harm and political divides (Nigatu and Raji 2024; Samuels 2020; Milmo 2021). Moreover, flawed moderation systems often censor benign content in non-English languages and silence marginalized voices (Elswah 2024b).

Regardless of the content modality, tech companies heavily rely on text-based methods, such as analyzing captions, transcripts, and subtitles, to automatically moderate harmful content (stream 2025a,b). However, this approach poses challenges for many languages spoken in the Global South, which are historically considered *"low-resourced"* due to the lack of high-quality datasets needed to train AI models (Rowe 2022; Nicholas and Bhatia 2023; Nigatu et al. 2024). However, data scarcity tells only part of the story. Economic and political oppression, insufficient human expertise, and limited access to digital infrastructures further exacerbate the *"low-resourcedness"* of these languages (Nigatu et al. 2024). Moreover, framing the problem solely as one of data scarcity overlooks broader challenges across the moderation pipeline, such as annotation, model training, and deployment. To address this critical gap, we examine the systemic barriers hindering equitable moderation for low-resource languages and explore actionable pathways to improve these systems. Specifically, we ask:

- **RQ1:** What systemic barriers impact automated moderation pipelines for low-resource languages?
- **RQ2:** How might we improve automated moderation for low-resource languages?

To address these questions, we conducted semi-structured interviews with 22 AI researchers and practitioners, specializing in harmful content detection and developing automated tools for diverse low-resource languages, such as Tamil from South Asia, Swahili in East Africa, Maghrebi Arabic from North Africa, and Quechua in South America.

Our findings reveal a spectrum of systemic issues beyond data scarcity impacting the automated moderation pipeline for low-resource languages. Many participants criticized tech companies' data restriction policy for hindering moderation research in the Global South. They pointed out that company's use of biased machine translation systems, Western-centric toxicity models, and poor language detection tools—overlook the cultural nuances of online harms and language evolution in the Global South. They emphasized that even with more data, current English-centric design of preprocessing techniques (e.g., tokenization, stemming) and language models disregard the linguistic diversity, morphological complexity, and dynamic evolution of languages through code-mixing and code-switching, which are often absent in English. For instance, unlike English which has a relatively fixed word order (Bender 2009), Tamil, Swahili, Arabic, and Quechua have agglutinative property, meaning they can form thousands of complex words from a single root. Data-driven models primarily trained on English typically fail to infer these linguistic properties that do not exist in English. As a result, words that frequently appear in sexual harassment, such as Tamil word *Mualichhu* (meaning, n**ples) incorrectly gets stemmed to *Mulai-* (meaning, sprout) and goes undetected by models.

Drawing on these findings, we use coloniality as a lens to critically examine how tech companies perpetuate digital colonialism (Kwet 2019), prioritizing profit over user safety in less profitable Global South markets (Nicholas and Bhatia 2023). These companies not only monopolize data extracted from the next billion users (Coleman 2018; Couldry and Mejias 2019), but also rely on biased sources for moderation—reinforcing harmful narratives about formerly colonized populations. We highlight how the English-centric, one-size-fits-all design of moderation tools reflects a colonial impulse by ignoring the linguistic diversity of Global South languages. Improving moderation for these languages requires more than technical fixes, as competing stakeholder priorities demand deeper systemic change. The key contributions of our work are as follows:

- A qualitative study using coloniality as a lens to examine how historical power imbalances skew automated moderation for low-resource languages in the Global South.
- A guideline to improve moderation for low-resource languages, acknowledging the practical and systemic issues.

2 Related Work

2.1 Content Moderation in the Global South

Content moderation refers to reviewing user-generated content to see if it aligns with tech company's policies on what should be allowed on the platform. Most companies use a mix of manual human reviews and automated AI models (Gorwa 2019). However, they often lack financial incentives to invest in moderation resources for less profitable markets in the Global South (De Gregorio and Stremlau 2023; Nicholas and Bhatia 2023). For instance, Meta funnels 87% of its global misinformation budget to the United States (US), despite Americans comprising only 10% of its user base (Popli 2021). The disparity is more glaring when companies swiftly handle harmful content from European countries that either offer strong economic incentives or hold geopolitical interest for the US (e.g., Russia-Ukraine war) (De Gregorio and Stremlau 2023). In contrast, tech companies have been less proactive in countering harmful content festering in many Global South countries (Milmo 2021; Wong and Ernst 2021; Wong and Harding 2021; Yibeltal and Muia 2023), while unjustly removing culturally and politically legitimate content from this region (Elswah 2024b; Shahid and Vashistha 2023).

Prior research links poor moderation in the Global South to tech companies' reliance on automated systems primarily trained on English and the lack of data in low-resource languages (Nicholas and Bhatia 2023; De Gregorio and Stremlau 2023; Nigatu et al. 2024). In contrast, little focus is given to other critical stages of moderation pipelines, such as who annotates what is harmful or what assumptions are made about deploying these models in complex, low-resource environments. To fill this critical gap, we explore the systemic challenges AI experts face throughout the moderation pipeline, placing these issues in the context of historical and colonial marginalization of non-English languages. We now present scholarly work critically examining power and control in content moderation systems.

2.2 Coloniality in Content Moderation

Decolonial scholars argue that colonial power structures persist by exploiting the labor and resources of formerly colonized populations while maintaining Western dominance over governance and knowledge production (Quijano 2000, 2007a). Decolonial computing challenges who designs technology, where it's developed, and how it impacts marginalized communities (Ali 2016). Whereas, postcolonial scholars critique the imposition of Western-centric technologies on diverse cultural contexts (Irani et al. 2010). Scholars have applied these lenses to examine content moderation systems.

Shahid and Vashistha (2023) used decoloniality to show how tech companies often impose Western values as global community standards, ignoring local socio-cultural norms when assessing online harms in the Global South. They draw parallels between Western-centrism in community guidelines and colonial suppression of Indigenous and marginalized communities' diverse ways of being, imposing Euro-modern rationality as the only legitimate way (Said 2000; Gramsci 2020; Quijano 2007a). Similarly, Siapera (2022) offers a decolonial critique of how tech companies dismiss input from racialized users when shaping policies against racist hate speech. She argues the race-blind, *"neutral"* moderation technologies mirror colonial legacies, criminalizing the identities and lived experiences of racialized communities as inferior (Quijano 2007b; Benjamin 2023).

In addition, scholars highlight how tech companies build AI moderation systems exploiting the labor and trauma of low-wage moderators from the Global South (Siapera 2022; Shahid and Vashistha 2023; Elswah 2024a). Tech companies often treat these moderators as disposable, concealing the nature of the work during recruitment, and evading responsibility for the harms moderators face (Ahmad and Krzywdzinski 2022; Elswah 2024a).

Moreover, moderation errors disproportionately harm marginalized communities, whose voices have long been silenced. AI models often reinforce systemic racism and heteronormative patriarchy by mislabeling Black and queer vernacular as toxic (Bhattacharyya 2018; Sap et al. 2019; Mohamed, Png, and Isaac 2020). Shaped by Western norms, these systems frequently misclassify culturally appropriate content in non-Western contexts as harmful while overlooking actual harmful content (Shahid and Vashistha 2023). For instance, Google's Perspective API underestimates toxicity in Swahili and Hindi but rates similar content in English and German more accurately (Udupa, Maronikolakis, and Wisiorek 2023). Udupa, Maronikolakis, and Wisiorek (2023) argues these biases persist because moderation systems inherit Eurocentric, colonial frameworks that justify unequal resource allocation across languages and regions.

We contribute to existing literature by critically examining the often-overlooked socio-political dynamics embedded across the automated moderation pipeline—from data collection to labeling, training, and evaluation. Since many studies overlook historical contexts when discussing power asymmetries in AI (Ovalle et al. 2023), we address this gap by interrogating the normative assumptions and design paradigms behind AI-driven moderation through the lenses of power and control. Drawing on interviews with AI experts, we explore how current language technologies shape moderation in diverse Global South contexts (RQ1) and how these practices might be improved (RQ2).

3 Methodology

To examine disparities in automated moderation pipelines, we interviewed 22 AI experts specializing in diverse low-resource languages spoken across the Global South.

Low-Resource Languages. We selected four diverse languages from different parts of the Global South. These are: Tamil (South Asia), Swahili (East Africa), Maghrebi Arabic (North Africa), and Quechua (South America) (see Table **??**). All these languages are considered low-resourced, despite having millions of speakers. UNESCO even declared Quechua as a vulnerable language due to systemic discrimination against Indigenous Quechua speakers in South America (Bank 2014). Due to limited resources, moderation errors are typically high for these languages. For instance, tech companies have repeatedly failed to address ethnic hate speech in Swahili (Witness 2022) and harmful content in Arabic (Elswah 2024b), while unjustly removing Tamil news articles as dangerous speech (Biddle 2022) and shadowbanning Arabic content on Palestine (Elswah 2024b).

Participants. We recruited people, who either (1) worked on automatic detection of harmful content, or (2) developed language models and tools in Tamil, Swahili, Maghrebi Arabic, or Quechua. We used purposive and snowball sampling to recruit 22 participants (see Table **??**). Among them, six specialized in Tamil, six in Swahili, five in Maghrebi Arabic, and three in Quechua. Most of them (n=15) were native speakers of one of these languages. Many of our participants were affiliated with academia (n=12) and trust and safety teams at Meta, OpenAI, and TikTok (n=4). Some worked for trust and safety vendors, who built moderation tools and datasets for different clients (n=2) and local AI startups (n=5). Some participants held multiple roles. Five self-identified as women and the rest as men. All participants had experience living in the Global South, such as Kenya, Tanzania, India, Sri Lanka, Peru, Morocco, and Egypt. Half of them were affiliated with Western institutions and were based in North America and Europe during the interview.

Data Collection and Analysis. We conducted 40-60 minutes long semi-structured interviews with the participants via Zoom. The semi-structured interviews focused on data collection, annotation, preprocessing, and model development in low-resource languages. We asked about the reliability and performance of models and tools they used to detect harmful content. The participants also reflected on biases and challenges they encounter throughout the process and discussed ways to address them. After each interview, we iteratively refined our interview protocol, stopping when the responses reached saturation. After obtaining ethical approvals from IRB, we conducted the interviews in English and audio recorded with the consent of participants. We compensated the participants with $100 Visa gift cards.

We transcribed the interviews, performed iterative open coding following reflexive thematic analysis (Braun and Clarke 2006), and continuously refined the emerging themes. Our coding process resulted in 441 codes, iteratively merged into 23 subthemes (e.g., model performance, annotation challenges)–which we mapped into different stages of automated moderation pipelines.

4 Findings

In this section, we outline systemic issues in moderating content in low-resource languages throughout automated moderation pipeline: data curation (4.1), annotation (4.2), preprocessing (4.3), and model training (4.4).

4.1 Barriers to Access Data on Harmful Content

To detect harmful content, most participants relied on large volumes of user-generated data from platforms like Facebook, X, YouTube, and Reddit. Industry practitioners reported easy access to such data within their companies, while academic researchers pointed to structural barriers, including lack of public datasets in low-resource languages and restricted access to social media data. For instance, during 2018–19, Twitter's free API limited researchers access to data older than two weeks. P14, a Swahili-focused academic researcher, commented:

> *"People frequently used the word 'madoadoa' [spots] to spew hatred and violence during the 2007-08 Kenyan election. But that changed in the 2022 election. Bad actors appropriated the popular song 'sipangwi' [I am not told what to do] and its plural form 'hatupagwingwi' to spread hatred. Unfortunately we neither have access to recent nor past data to study how hate speech tactics have evolved."*

To bypass API restrictions, many researchers and small trust and safety vendors used open-source scrapers to collect historical user-generated content. However, these tools often misclassified romanized or code-mixed content as English due to the use of Latin scripts. Recently, companies like Meta, X, and Reddit have blocked scrapers and restricted API access to user content (e.g., Meta's shutdown of CrowdTangle) (Mehta 2023; Bellan 2024; Stokel-Walker 2024; Perez 2024). These restrictions, combined with limited datasets, have severely hindered efforts to study disinformation and hate speech in high-risk Global South regions. P18, a trust and safety practitioner at TikTok remarked:

> *"After ChatGPT came out, companies are cautious of publicly sharing their data given the competition to develop their own language models. That's why we no longer see that openness around sharing data."*

Due to challenges in accessing and curating datasets, some researchers ceased studying online harms in their regions. Others started collecting online posts manually, using data from shared tasks at NLP conferences, deploying

	Tamil	Swahili	Maghrebi Arabic	Quechua
Number of speakers	80 million	100 million	88 million	8 million
Geographic region	South Asia: Tamil Nadu (India), Sri Lanka, etc.	East Africa: Kenya, Tanzania, etc.	North Africa: Morocco, Algeria, Tunisia, etc.	Andes: Bolivia, Peru, Ecuador, etc.
Language family	Dravidian	Bantu	Semitic	Quechuan
Grammar	Agglutinative, subject-object-verb (SOV)	Agglutinative, subject-verb-object (SVO)	Root based, verb-subject-object (VSO)	Agglutinative, subject-object-verb (SOV)
Colonial influence	British	Portugese, German, British, Arabic	French, Spanish, Italian	Spanish

Table 1: Various characteristics of four low-resource languages featured in this study.

ID	Language expertise	Role	ID	Language expertise	Role
P1	Tamil	Professor, Startup founder in India	P12	Swahili	Professor in Kenya
P2	Tamil	Professor in India	P13	Swahili	Master's student in Kenya
P3	Tamil	PhD Student in Europe	P14	Swahili	Professor in Tanzania
P4	Tamil	Master's student in US	P15	Swahili	Startup founder in UK
P5	Tamil	Software engineer in India	P16	Swahili	ML engineer in Kenya
P6	Tamil	Lecturer in Sri Lanka	P17	Swahili	Industry practitioner at Meta
P7	Indic languages	Trust and safety vendor	P18	Arabic	Industry practitioner at TikTok
P8	Indic languages	Trust and safety vendor	P19	Arabic	Industry practitioner at Meta, OpenAI
P9	Quechua	Linguist in Europe	P20	Arabic	Industry practitioner at Meta
P10	Quechua	PhD student in Europe	P21	Arabic	PhD student in UK
P11	Quechua	Lecturer in Peru	P22	Arabic	Startup founder in Europe

Table 2: Demographics of participants in our study.

community surveys, or voluntary WhatsApp data donations. These methods were time-consuming and yielded small, inconsistent datasets, often insufficient for training AI models.

Several participants emphasized the need for equitable access to user-generated content. While acknowledging privacy concerns, they criticized platforms like TikTok for limiting API access to researchers in the US and Europe (TikTok 2024). In response to such disparities, African researchers have joined grassroots initiatives like Masakhane and Tanzanian AI to reclaim ownership of locally generated data. Academics also condemned tech companies for mishandling harmful content in their regions while monopolizing user data. P19, who worked at Meta commented:

> *"When I worked at Meta, the trust and safety team prioritized the US. These for-profit corporations derive most of their revenues from Western markets. Although Europe has strong regulatory policies, those markets are important to the company. So the prioritization simply reflects that."*

Participants raised concerns about how tech companies address data scarcity in low-resource languages. They criticized keyword-based filtering for overlooking dialectal variation and treating these languages as monolithic. They also flagged biases in machine-translated texts, often used as a workaround for limited data. For example, Kenyan Swahili researchers noted that Google Translate heavily supports outdated Sheng (a cerole combining Swahili and English) over its modern variants like Shembeteng, while Tanzanian researchers found it biased toward Kenyan Sheng over standard Tanzanian Swahili. Similarly, Quechua researchers highlighted tech company's problematic reliance on outdated Bible translations and colonial-era texts. One industry practitioner also noted their company using old Arabic dictionary due to scarce datasets in Maghrebi Arabic. Participants emphasized that models built on outdated corpora and biased machine translations are ill-suited to address the evolving nature of hate speech online.

Small AI startups and trust and safety vendors reported that big tech companies often showed interest in their tools and datasets–only if offered for free. They demanded tech companies to support grassroots, local research efforts to address online harms in the Global South. P9, a Quechua linguist expressed:

> *"They [companies] should work with us, indigenous Quechua people, to build corpuses instead of taking the shortcut by using machine-translated texts. We found rule-based translation that incorporates grammatical knowledge works better for Quechua than stochastic methods, which require lots of data that do not exist in Quechua. When we contacted Google, they proposed us to work voluntarily. But I worry they will try to appropriate our free labor."*

These findings show that tech companies' lack of support for local data curation efforts and gatekeeping of user-generated content worsen data scarcity, disrupting grassroots efforts to tackle online harms in the Global South.

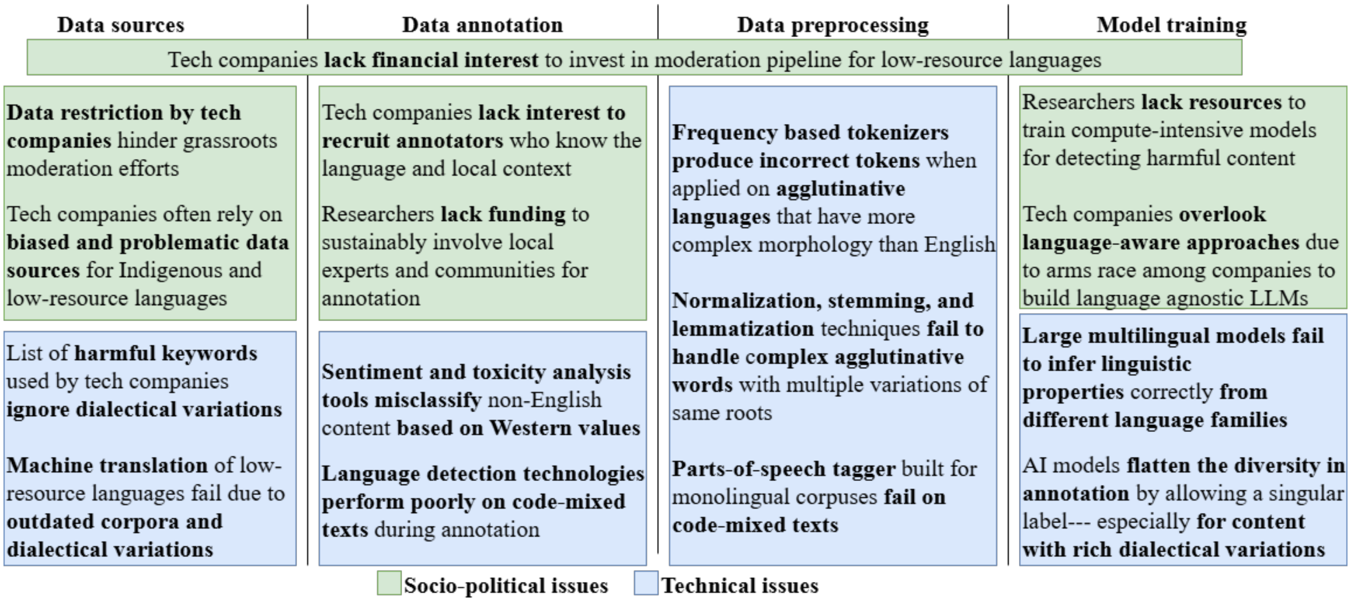

Figure 1: Issues affecting different stages of automated moderation pipeline for low-resource languages.

4.2 Difficulties in Annotating Harmful Content

Annotation involves labeling the data to train AI models to detect harmful content and identify its type. Tech companies frequently outsource annotation to human moderators in the Global South. One industry practitioner noted that their company assigned Kenyan moderators to annotate Swahili dialects they didn't understand. Their efforts to assign content to moderators with appropriate language expertise often fail due to poor language identification tools for low-resource languages. Participants stressed that companies have always underfunded annotation efforts for languages spoken in *"less profitable regions."* P19, who worked at a social media company, shared:

> *"During Arab Spring, [redacted] had only two Arabic speaking moderators. There's so much diversity in the Arab world– it's unlikely that the two moderators will get the full context of Arab Spring in Tunisia or Green Movement in Iran. Although a lot has changed since then, the core structure and issues remain the same."*

Participants expressed that tech companies' lack of understanding of ground realities, socio-cultural norms, and linguistic nuances–significantly hinder their ability to address harmful content in the Global South. They stressed that tech companies should *"give Global South a seat at the table"* when annotating hate speech. P12, a Kenyan researcher working on Swahili, remarked:

> *"It matters who is defining hate speech. We noticed that people use superlatives and 'US' vs. 'Them' narrative to spread supremacist views and ethnic hate speech. We developed our annotation framework to capture these cases. Since Twitter did not remove these tweets, their definition of hate speech must be different. By allowing these posts Twitter is reinforcing stereotypes about Africans being violent."*

To ensure that the annotated datasets capture local sensitivities, local researchers often involved linguists, activists, and affected communities to inform their annotation guidelines. P3, a researcher working on Tamil, explained:

> *"It's important to consider intersectionality when annotating hate speech in multicultural environments like India, where caste, religion, and gender are intertwined. For example, we found 'shuttlecock' [badminton cork] is used as a derogatory term against Muslim women who wear burka. Our team of feminist activists, experts on gender studies, and survivors of harassment helped us annotate coded hate speech that are both misogynist and Islamophobic. Similarly, there were innocuous comments like 'you are my sweetheart.' When companies recruit gigworkers who are usually male, they would rate this as harmless. But since we worked with victims of sexual harassment and recipients of such comments, they could recognize these messages are part of broader harassment Indian women face online."*

Researchers valued involving community partners in data annotation but faced funding challenges that hindered sustained collaboration, annotator training, and quality control. Due to limited resources, they often relied on undergraduates to annotate hate speech and toxic content without being able to provide mental health support. P11, an academic researcher working on Quechua, shared:

> *"Very often the dataset we are creating is the first of its kind in Quechua. Although experts and community members are willing to help voluntarily, it's difficult to sustain their free labor in the long run to annotate large volumes of data. So, we often strategize to annotate only a subset of data. We can't rush people to annotate faster because they are helping out of gen-*

erosity. Thus, it takes months to annotate anything."

To make the most of limited annotation resources, researchers often used sentiment and toxicity analysis tools to find negative content, reducing the sample size for manual annotation. However, they noted that existing free and proprietary tools from tech companies often lack cultural nuances. P13, a researcher specializing in Swahili, elaborated:

> *"In America, people casually use the word 'dawg' to refer to buddy but in Kenya calling someone dawg will be disrespectful. Similarly, in America people think calling 'fat' is body shaming. In Africa fat is considered beautiful and opulent. But Google's perspective API missed these cases by applying American scale."*

Tamil researchers shared that they lose valuable annotation time and budget when manually verifying target languages in scraped corpuses because language identification tools have poor coverage for most low-resource languages. These tools often fail to separate code-mixed Tanglish (Tamil-English) from Kannada-English or Telugu-English because Tamil, Kannada, and Telugu often share words with same roots. Similarly, Maghrebi Arabic researchers reported that these AI tools often fail to differentiate between Arabic, Farsi, and Urdu due to overlapping scripts.

These findings reveal that limited AI support for low-resource, non-English languages hampers annotation efforts, which are further strained by chronic underfunding of grassroots initiatives. Despite having ample resources, tech companies frequently overlook cultural nuances of online harms due to insufficient engagement with stakeholders and communities in the Global South.

4.3 Preprocessing Challenges for Harmful Content Detection

Preprocessing involves cleaning and transforming raw data in a suitable format to train AI models. Our participants faced several challenges when applying existing preprocessing techniques on low-resource languages.

Tokenization. Tokenization involves segmenting text into words or subwords for language processing. Participants noted that multilingual models like BERT and RoBERTa use frequency-based tokenization methods (e.g., WordPiece, BPE), generating tokens based on the frequency of words or co-occurring character pairs. These methods, primarily designed for English, underperform on languages like Tamil, Swahili, Maghrebi Arabic, and Quechua, which have more complex morphology than English. They explained that these languages have agglutinative properties, forming complex words by combining multiple morphemes (i.e., the smallest unit of meaning), where each morpheme retains its original meaning. For example, the Quechua word *'rimanqakuma'* (meaning, they will definitely speak) consists of three morphemes: *'rima-'* (meaning, to speak), *'-nqa'* (refers to future tense) and *'-kuma'* (signifies emphasis). The final meaning is directly derived from these constituent morphemes. P9 elaborated further stressing the need to derive morphemes correctly during tokenization:

> *"Frequency based tokenizers have been designed considering English as a model language. Since English is data-rich, frequency based method really works well. But for low-resource, agglutinative languages it creates illegible tokens by wrongly splitting the morphemes. If we train models with wrongly split tokens, the models won't derive correct embeddings. Instead, when we used linguistically motivated tokenizer, the performance significantly improved for Quechua in downstream tasks."*

Maghrebi Arabic NLP researchers also noted that using specialized morphological and monolingual tokenizers improve sentiment analyses for diverse low-resource languages, typically underrepresented in multilingual models. Swahili researchers further highlighted the challenges of tokenizing code-mixed hashtags that are often used to incite attacks while evading detection by platforms. For example, in the Sheng hashtag *#TupataneTuesday* (meaning, let's meet each other on Tuesday), used by protesters, the Swahili word *Tupatane* must be correctly segmented into its morphemes: Tu- (we), -pat (to meet), -ane (each other). However, poor performance of language identification technologies on code-mixed texts complicates the application of tokenization algorithm based on language.

Normalization. Researchers also identified challenges in the normalization process, where words are converted to their standard forms before tokenizing (e.g., baaaad is normalized to bad). Some participants reported that non-standard spelling of agglutinative words causes confusion during normalization. P6, a Tamil researcher from Sri Lanka, explained:

> *"In Tamil, 'Amma' means Mother and 'Ama' means Yes. On social media people often enthusiastically write Ama as 'Aammaa' (similar to Yeessss) or distort the word Amma as 'Aammaa' in gendered slurs. The model often makes errors while normalizing such cases and fails to flag offensive language."*

Stemming and Lemmatization. These steps are performed to reduce words to their meaningful roots before training models (e.g., beautiful and beautify are reduced to beauty). Several participants reported facing challenges because existing tools have higher error rates in complex agglutinative languages, where *"each root can take thousands of inflected forms"*, than morphologically simpler languages like English. P6 further described:

> *"Both understemming and overstemming of complex Tamil grammar can cause error in detecting offensive language. Words like Mulaicchu (meaning, n**ples) often wrongly gets stemmed to Mulai- (meaning, sprout) and then gets ignored by model."*

Parts-of-Speech (POS) Tagging. Some participants reported that since most models are trained on English, which is a subject-verb-object (SVO) language, it leads to errors on languages that follow subject-object-verb (SOV) structure. Therefore, they performed POS analysis during data preprocessing to give models additional contexts about derogatory adjectives and verbs aimed at individuals or groups (nouns).

For example, in *Nāyai seruppāla aṭikkaṇum* (meaning, beat the dog with sandals) the object (noun) *Nāyai* appears before the verb *aṭikkaṇum*. However, researchers faced several challenges in detecting POS due to code-mixing. P2, an academic researcher, shared:

> *"When I started doing NLP research in early 2000, there was no POS tagger for Tamil. There was barely any dataset to work with. We built corpuses from scratch and worked with linguists to annotate complex Tamil vocabulary. But the POS tagger based on monolingual Tamil does not work well on Tanglish from social media. Although many frame code-mixed data as problematic and low-quality, this is the reality of how social media users from non-English speaking countries write online. Handling code-mixing is very challenging. But we don't have access to code-mixed data from social media since they stopped access."*

These findings show that current preprocessing techniques, predominantly developed with English in mind, do not account for the morphologically rich and code-mixed nature of languages in the Global South, reflecting historical imbalances in linguistic and technological priorities.

4.4 Challenges in Developing AI Models for Harmful Content Detection

After standard preprocessing, the data is fed into AI models for training and detecting harmful content. Participants reported using multilingual models, such as Google's mBERT, Facebook's XLM-RoBERTa, and AI4Bharat's IndicBERT but found these models perform poorly when detecting harmful content in low-resource languages. They explained that although these data-driven models are designed to be language-agnostic, being primarily trained on high-resource languages like English, they better learn the simpler morphology and fixed word orders of English. In contrast, data sparsity in low-resource languages limits these models' ability to capture the rich inflectional morphology, agglutinative property, complex grammar, and diverse word orders that are absent in English. P4, specializing in Tamil, described:

> *"English and Tamil are from different language families and Tamil has richer morphology than English. How can these models derive correct embeddings of complex Tamil words by computing from the point of view of English? That's why IndicBERT doesn't perform well. There, Hindi and Marathi are from the same family but Tamil is a Dravidian language. So without considering the specifics of language families, you can't get performance improvement."*

Researchers cautioned that adding data from multiple languages can degrade model's performance in both low-resource and high-resource languages due to limited model capacity, a phenomenon known as the *"curse of multilinguality"* (Chang et al. 2023). Additionally, they criticized how AI models cannot handle diversity in annotations, especially for languages like Tamil and Swahili that have tens of dialectical variations. P15, a startup founder focusing on Swahili AI, explained:

> *"In Swahili the word 'right' has at least 20 different transliterations depending on the context. Similarly, in my region, the word 'Mathikkalla' refers to 'I could not recognize you' but in other regions, the same word means 'to neglect someone.' So, annotators would label the same content differently depending on their region. This impacts offensive language detection because current approaches to train AI models flatten annotator diversity into a singular label."*

Some researchers observed that large language models frequently misclassify code-mixed content during hate speech detection, especially when the spelling and words signal non-Western ethnicity. Trust and safety practitioners attributed these errors to a lack of diversity within tech companies and shared that very often their teams are linguistically and culturally homogeneous. They commented that company's diversity efforts often end at recruitment; once hired, employees have to work following company's priorities, which are typically centered around English. P20 remarked how this lack of diversity leads to biased models:

> *"In Western media, Arabic phrases, such as 'Allahu Akbar' [God is great] mostly appear in the context of terrorism. When companies train AI models on such articles, the models learn these negative associations. But there is none in these teams to inform that local people use these phrases to express everyday joy and sorrow, beyond the instances of extreme speech portrayed by Western media."*

Industry practitioners shared that despite the shortcomings of large language models in low-resource languages, their companies are prioritizing AI models over alternative linguistic approaches they used in the past. They emphasized the advantages of using AI models, particularly in reducing the burden of tedious and distressing moderation work for humans. In contrast, AI researchers and practitioners working in the Global South highlighted their struggle in training billion parameter models due to a lack of funding, computational power, and appropriate hardware. For example, Swahili researchers and engineers shared that they could not buy GPUs in Kenya and Tanzania and had to rely on their contacts in the US to access these resources. Many pointed out that free resources from Google Colab and Kaggle are barely enough to experiment with, train, and deploy these large language models. P12 commented:

> *"We lack the necessary data, funding, and resources to build dedicated models for our languages. We spend time scraping for little data and cleaning it. I hope we can decolonize NLP research on online harms, so that we don't have to rely on technologies favoring high-resource languages like English and developed for nations with lots of computing power."*

These findings highlight that the resource-intensive and English-centric design of large language models is ill-equipped to address online harms in low-resource languages from the Global South, reflecting how the needs of these communities are usually sidelined in the development of AI-driven moderation technologies.

5 Discussion

While prior work attributes moderation challenges in low-resource languages to the lack of labeled datasets (Rowe 2022; Nicholas and Bhatia 2023), our study uncovers how socio-political factors in technology design exacerbate these issues. Our findings underscore how tech companies continue to rely on biased machine translation systems using outdated corpora instead of collaborating with experts and communities from the Global South—often appropriating their free labor. We reveal how these companies' blanket data restrictions for building proprietary large language models aggravate data scarcity to address online harms in these regions (4.1). While prior studies report biased and opaque annotation practices among tech workers (Scheuerman and Brubaker 2024) and ML researchers (Geiger et al. 2020), we examine the structural factors enabling these issues. Socio-political issues, such as tech companies' weak financial incentives to improve annotation for Global South languages and limited funding available to Global South researchers along with technical issues like Western-centrism in sentiment and toxicity models and treating code-mixed data as "poorer quality" when developing language detection tools–compromise the annotation process (4.2).

Moreover, most studies explain away preprocessing and model building challenges in low-resource languages by highlighting data scarcity (Khan et al. 2023; Zhong et al. 2024). In contrast, our study questions the status quo that prioritizes data-intensive methods while overlooking alternative approaches that center linguistic diversity, morphological complexity, and dynamic evolution through code-mixing and code-switching—phenomena largely absent in English (4.3, 4.4). We provide concrete examples of how normative assumptions in technology design contribute to moderation errors in diverse Global South languages–that remain invisible when assessed solely through low accuracy rates. Our focus on diverse languages help us establish the systemic nature of moderation biases. In discussion, we probe deeper into these systemic inequities, unpacking their historical and socio-political roots–often overlooked in existing discourse (5.1). We then discuss approaches to improve moderation for low-resource languages while acknowledging the complexity of the issue (5.2).

5.1 Coloniality in Moderation Pipelines

Data Curation. Our findings show that tech companies lack interest to expend moderation resources for less profitable markets in the Global South. Our participants stressed that companies benefit by monopolizing user-generated data to train proprietary large language models, while restricting researchers' access to the very data needed for detecting harmful content. For instance, shortly after Reddit locked public data (Perez 2024), it partnered with OpenAI to enable training ChatGPT on its content (OpenAI 2024). Similarly, Meta launched AI across Facebook, WhatsApp, and Instagram to train proprietary models on public posts without letting users opt out (Jiménez 2024), while simultaneously closing CrowdTangle that allowed researchers to access public content on Meta (Bellan 2024). Researchers criticized these blanket restrictions on public data as privacy washing, impeding trust and safety scholarship within academia and civil society (Arney 2024).

These restrictions disproportionately affect researchers and practitioners in the Global South, where datasets in non-English languages remain scarce. This data scarcity stems from colonial legacy that suppressed Indigenous and native languages in the Global South (Thiong'o 1986; Bank 2014; Obi-Young 2018; Kolli 2024) and deprioritized their digitization and technology development (Bird 2020; Schwartz 2022; Held et al. 2023; Ògúnrẹ̀mí, Nekoto, and Samuel 2023). The systemic omission affects all downstream NLP tasks in low-resource languages, including automated moderation– further hampered by data restriction imposed by tech companies.

Our participants highlighted that the data controlled by tech companies are generated through the unpaid labor of users in their communities. Coleman (2018) explains that Facebook introduced Free Basics initiatives in the Global South to extract data from the region's next billion users, taking advantage of weak data protection laws and regulatory frameworks. Kwet (2019) likens this process to digital colonialism. He argues that much like colonizers who built railroads to extract material resources from colonies, tech companies control digital infrastructures in the Global South, reduce local communities to products rather than producers, and commodify their data for corporate profit.

Our analysis reveals that tech companies' reliance on cheap web-scraped data, machine translations, and religious texts for low-resource languages (Kreutzer et al. 2022; Christodouloupoulos and Steedman 2015; Ghosh and Caliskan 2023)–introduces significant biases in moderation. This includes wrongly associating Arabic phrases with terrorism and normalizing extreme speech in African contexts. Such biases reflect digital orientalism (Alimardani and Elswah 2021), where colonial perspectives shape discriminatory narratives regarding the colonized 'other' (Said 1977). Likewise, the use of colonial-era texts to build Quechua datasets overlooks the historical role of colonial churches in suppressing Indigenous languages, while appropriating them only for cultural control (Heller and McElhinny 2017, p. 29). Thus, our findings highlight how colonial legacies continue to shape the availability of data sources–required to study online harms in the Global South.

Annotation. Trust and safety practitioners in our study noted that tech companies lack incentives to hire expert moderators for Global South content. Yet, these companies often outsource English-language annotation to the Global South, exploiting low wages and weak labor protections, and inflicting psychological harm to local moderators (Elizabeth Dwoskin and Cabato 2019; Elswah 2024a). This practice mirrors colonial exploitation, where the Global South workforce serves the interests of the Global North with little regard for local needs or safety (Posada 2021; Malik 2022).

Additionally, we found that limited funding in Global South institutions hinders grassroots efforts to annotate harmful content in local languages. Historically, resources extracted through colonial exploitation enabled Western nations to advance their scientific agenda and build extensive datasets (Schöpf 2020). Consequently, most misinformation research focuses on the West due to easy availability of annotated datasets in English (Rananga et al. 2024). These systemic inequities, marked by resource scarcity in the Global South and tech companies' disinterest in investing in these regions (Nicholas and Bhatia 2023; De Gregorio and Stremlau 2023)–further limit the availability of annotated datasets in low-resource languages.

NLP Tools Used in Moderation. Our findings underscore that current NLP technologies, primarily designed for English, overlook the cultural context, linguistic complexity, and evolution of languages in the Global South. For example, our participants reported that Google's Perspective API misinterprets diverse notions of toxicity across different cultures. Similarly, Das et al. (2024) demonstrate that sentiment analysis tools for low-resource languages disproportionately associate negative sentiment with certain religious and national identities– replicating colonial hierarchies of division sowed by British rulers in the Indian subcontinent.

Decolonial scholars and historians have long documented the colonial project of standardizing European languages by creating dictionaries and grammars to assimilate Indigenous populations while suppressing local languages (Fishman 1989; Heller and McElhinny 2017; Anderson 2020; Fanon 2023). These forced affected communities to code-switch between native and European languages to navigate colonized spaces (Mufwene 2004). These legacies resulted in poor early support for non-Latin scripts online, continuing to hinder participation from speakers of many low-resource languages (van Esch et al. 2019; Held et al. 2023; Nigatu et al. 2024). This discrimination has forced non-English speakers to adopt romanization and code-mixing for communicating online (Held et al. 2023). However, the closed, proprietary language models, relying on *sanitized* datasets, disenfranchise local knowledge, impose Western normative values without empowering local communities to align the model to their own values, forestall alternative visions, and perpetuate colonial binaries that frame advanced technologies as rescuing "primitive" languages (Verran and Christie 2007; Bird 2020; Varshney 2024).

Primarily being trained on English, these language models perform well on languages that share important typological properties with English (Bender 2009; Arnett and Bergen 2024). Thus, these models fail to capture the elaborate morphology present in many low-resource languages. Historically, linguists considered agglutinative languages as *"less evolved"* than Western languages, such as Spanish, Greek, or German (Errington 2007). Bender (2009) critiques AI models for making assumptions about language structures that advantage some languages at the expense of others, highlighting their inherent lack of language independence. Scholars criticize such one-size-fits-all solutions for embodying *"colonial impulse"* that disregards the ecology of diverse languages and perpetuates colonial hierarchies (Dourish and Mainwaring 2012; Bird 2022). For languages spoken in the Global South, this translates to collapsing their linguistic diversity and complexity to a simplistic construct of data scarcity– often taken at the face value.

In sum, our findings show that existing challenges affecting automatic detection of harmful content in low-resource languages are often systemic and run deeper than the mere availability of data.

5.2 Considerations for A Path Forward

Tackling harmful content in low-resource languages is a complex issue shaped by conflicting priorities across stakeholders. Tech companies often see investing in moderation systems for these languages as *unprofitable* despite these languages having millions of speakers (Nicholas and Bhatia 2023; De Gregorio and Stremlau 2023). The deprioritization of trust and safety efforts within US-based tech companies undermines global accountability, reinforcing a US-centric vision of free speech (Scarcella 2024; Divon and Ong 2025). Researchers also face little incentive to build labeled datasets for low-resource languages due to high effort (Sambasivan et al. 2021) and limited academic payoffs (Held et al. 2023). Governments in many Global South countries often resort to censorship or data localization laws due to platforms' failure to address harmful content but end up facing pushback from US tech lobbies (Kak 2020; De Gregorio and Stremlau 2023). Meanwhile, civil society groups in the Global South frequently feel marginalized as tech companies often treat local collaboration as a checkbox exercise rather than genuine partnership (Centre 2024). Acknowledging these realities, we outline some concrete steps to make content moderation more equitable.

Strengthening Local Research Capacity. Prior research highlights that when Global North institutions are funded to develop models for low-resource languages without involving local experts, they often fail in context-specific moderation tasks (Nicholas and Bhatia 2023). Bhabha (2011) argues that enhancing national resources of the Global South is essential to addressing the geo-politics of resource distribution and facilitate redistributive justice. Therefore, governments, grant-making agencies, and research award programs by tech companies must invest in building self-sustaining, grassroots research ecosystems that actively engage local experts from the Global South. For example, the AI4D Africa program, funded by international governments and research institutes, supports the development of local AI research hubs and talent, empowering African researchers to lead projects that address their communities' needs (IDRC 2024). Initiatives like Masakhane in Africa, AI4Bharat in India, and ARBML in the Arab World, which are democratizing AI research on low-resource languages, should be strengthened through targeted funding to amplify their impact.

Labeled Datasets. Although social media companies frequently cite privacy concerns in data sharing, established practices from other fields suggest feasible solutions. For instance, the Yale Open Data Access Project allows medical companies to securely share anonymized clinical trial data with vetted researchers for approved studies (Nicholas and Thakur 2022). Similarly, differential privacy techniques can protect personal information when sharing large datasets (Kapelke 2020; Garfinkel and Bowen 2022). Since for-profit companies have little incentive to share user data, policy measures like the Digital Services Act (DSA Articles 12 and 40), which grant vetted researchers access to data from very large online platforms (Commission 2025), can offer a potential solution. The DSA should seriously consider the data needs of researchers building automated tools to address online harms in low-resource languages.

Additionally, for languages with a significant digital presence, voluntary data donation by native speakers can be useful for grassroots researchers. For example, Garimella and Chauchard (2024) developed a data donation tool for closed WhatsApp groups while safeguarding the privacy of both donors and their contacts. In contrast, for Indigenous languages with limited digital presence, building respectful and equitable community relationships is essential, prioritizing local agency in community partnerships (Bird 2020; Thakur 2025). Indigenous Protocol and AI working group has created guidelines on how to integrate Indigenous ontology and epistemology when building AI tools (Lewis et al. 2020) that can educate AI practitioners to not rely on colonial epistemology (e.g., Bible data) for Indigenous languages.

Participants emphasized the need for diverse perspectives when annotating intersectional hate speech, urging companies to prioritize high recall to better capture harmful content (Parrish et al. 2024). For instance, during the Tigray conflict, moderators from dominant ethnic groups influenced how ethnic hate speech got labeled, while minority moderators–often at the bottom of the hierarchy–had limited influence on platform policy (Iyer 2025). Since for-profit motives often discourage companies from diversifying moderators, governments can apply regulatory pressure–similar to data localization laws–to require platforms to recruit local moderators for content from their respective regions. For example, India's Information Technology (IT) Rules, 2021 mandate major social media companies to appoint local grievance officers and compliance staff based in India, effectively pressuring companies to localize their moderation teams and workflows (MeitY 2023).

Moreover, it is important to follow the best practices for supporting community labor when annotating harmful content by disclosing the task, offering opt-out options, providing well-being support, and monetary compensation (Radiya-Dixit and Bogen 2024). For example, Karya—a nonprofit data company based in India—empowers disadvantaged communities through data annotation work and pays them nearly 20 times more than the local minimum wage (Perrigo 2023). Local labor rights organizations can partner with initiatives like the Global Trade Union Alliance of Content Moderators (Union 2025) to advocate for fair treatment of moderators. Together, they can pressure tech companies to balance workloads when assigning traumatizing content and to provide fair wages, mental health support, and safer working conditions. (Elswah 2024a).

Language-Aware Solutions. Since current NLP tools and language models are inherently English-centric, our participants recommended approaches that incorporate linguistic knowledge, such as using morphological segmenters instead of frequency-based tokenizers (Abdelali et al. 2021; Zevallos and Bel 2023), rule-based translations over stochastic machine translations (Sreelekha, Bhattacharyya, and Malathi 2018), and vector embeddings of local hateful phrases for detecting code-mixed hate speech (Devi, Kanimuthu, and Madasamy 2024). Some participants also discouraged using multilingual models for moderation given these models fail to infer correct linguistic context for different language families. However, given the arms race among tech companies to develop large multilingual models (Gupta 2024), it is unlikely that they will shift to such linguistically informed solutions without regulatory pressures. Meanwhile, limited access to computing power in the Global South limits researchers' capacity in training and experimenting with *"language-specific"* (monolingual) models and *"language-aware"* approaches that do not necessarily rely on vast datasets or huge computing power. Free computing resources provided by tech companies, such as Google's Colab and TPU Research Cloud programs–remain inadequate for research purposes. Expanding access to these resources is critical to enable more equitable and inclusive moderation research in Global South languages.

Policy and Practice. For the data and tools created by local experts to have meaningful impact, they must be deployed to moderate local content. Cohere's partnership with HausaNLP to integrate African language datasets into its multilingual Aya model (Radiya-Dixit and Bogen 2024) demonstrates the potential of such efforts. Similarly, the scales created by local researchers to evaluate model's performance for detecting code-mixed hate speech (Das et al. 2022)–should be integrated into tech companies' evaluation frameworks. Governments and civil society should create regulatory policies that require tech companies to prioritize local representation, data ownership, and community self-determination. These policies must articulate the specific harms caused by flawed content moderation systems in high-risk, under-resourced contexts, rather than simply applying Western frameworks of harm to regional trust and safety efforts (Kennedy and Campos 2025). Additionally, they should push for the inclusion of model performance metrics for low-resource languages in transparency reports, evaluated against locally defined benchmarks. Given the limitations of accuracy metrics in class-imbalanced content moderation task, regulatory frameworks should mandate the inclusion of more informative measures such as recall (% of correctly flagged harmful content) and precision (% of flagged content that is truly harmful) (Wei, Zufall, and Jia 2025). Such steps could surface the limitations of current models and incentivize progress toward better moderation of harmful content in underrepresented languages.

6 Conclusion

Content moderation is difficult but even more so for low-resource languages due to systemic issues and colonial legacies impacting data access, culturally-aware annotation, and the incentive to build language-aware technologies. As a result, the techno-solutionist approaches championed by Silicon Valley fail to address the multifaceted complexities of detecting online harms in low-resource languages. To break this vicious cycle, we need a bottom-up approach that invests in digital public infrastructure and prioritizes online safety over corporate profit.

7 Positionality Statement

All authors of this work come from historically colonized regions in the Global South and are native speakers of languages, which are considered *"low-resourced."* All authors have extensive experience of doing critical research in diverse Global South contexts. Although none of the authors are affiliated with industry, our background in academia (computer science and communications) and civil society—enabled us to engage with participants both at technical and policy levels. Although we come from historically colonized countries, we are affiliated with academic institutions that benefited from colonial expansion and were established using wealth derived from the forced appropriation of Indigenous lands. We acknowledge that these affiliations afford us research privileges, such as access to funding, institutional support, and global visibility that are often inaccessible to many Global South researchers. In this tension, we identify with Villenas (1996)'s notion of having *"feet in both worlds,"* as we simultaneously belong to communities shaped by colonial histories and to academic institutions that have profited from those same colonial legacies.

Similarly, we recognize that caste, religion, ethnicity, and other intersectional identity likely shaped the experiences of our participants although we did not explicitly collect such data. For instance, several Indian participants studying Tamil were likely from upper-caste backgrounds, some researchers studying Quechua were not of Indigenous origin, and many had greater access to resources due to Western affiliations compared to their counterparts in the Global South.

Following Fine (1994)'s self-reflexive approach, we acknowledge the intersections of privilege and marginalization among our participants. For instance, Quechua-speaking Indigenous researchers reported greater struggle in practicing their native language in academic spaces compared to Western researchers studying the same Indigenous language. Similarly, the experiences of upper-caste AI experts would differ from those coming from lower-caste backgrounds–who might face additional challenges and lack social capital in pursuing AI research in Tamil. Swahili researchers and practitioners from Tanzania felt their Kenyan counterparts enjoyed more visibility since most tech companies have regional offices within Kenya. Although industry practitioners had greater access to computing resources, they reflected on their limitations within corporate financial infrastructures and lack of diversity within these organizations. As Haraway (2013) argues, knowledge production is always situated; the experiences of AI researchers and practitioners in our study form a *"partial perspective"*–shaped by their backgrounds and the struggle of researching marginalized, low-resource languages.

Acknowledgments

We are earnestly thankful to Internet Society (ISOC), Dhanaraj Thakur, Aliya Bhatia, Cornell Global AI Initiative, Infosys, Justine Zhang, DeVan Hankerson Madrigal, Gabrierl Nicholas, and Sarah Zolad for supporting this work.

References

Abdelali, A.; Hassan, S.; Mubarak, H.; Darwish, K.; and Samih, Y. 2021. Pre-training bert on arabic tweets: Practical considerations.

Ahmad, S.; and Krzywdzinski, M. 2022. *Moderating in obscurity: How Indian content moderators work in global content moderation value chains*, chapter 5, 77–95. Cambridge, MA: The MIT Press.

Ali, S. M. 2016. A brief introduction to decolonial computing. *XRDS: Crossroads, The ACM Magazine for Students*, 22(4): 16–21.

Alimardani, M.; and Elswah, M. 2021. Digital orientalism:# SaveSheikhJarrah and Arabic content moderation.

Anderson, B. 2020. Imagined communities: Reflections on the origin and spread of nationalism. In *The new social theory reader*, 282–288. United Kingdom: Routledge.

Arnett, C.; and Bergen, B. K. 2024. Why do language models perform worse for morphologically complex languages?

Arney, J. 2024. Data dump: Meta killed CrowdTangle. What does it mean for researchers, reporters?

Bank, W. 2014. Discriminated against for speaking their own language.

Bellan, R. 2024. Meta axed CrowdTangle, a tool for tracking disinformation. Critics claim its replacement has just '1% of the features'.

Bender, E. M. 2009. Linguistically Naïve != Language Independent: Why NLP Needs Linguistic Typology. In *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?*, 26–32. Athens, Greece: Association for Computational Linguistics.

Benjamin, R. 2023. Race after technology. In *Social Theory Re-Wired*, 405–415. New York: Routledge.

Bhabha, H. K. 2011. *Our neighbours, ourselves: Contemporary reflections on survival*. De Gruyter.

Bhattacharyya, G. 2018. *Rethinking racial capitalism: Questions of reproduction and survival*. Maryland, USA: Rowman & Littlefield.

Biddle, S. 2022. Facebook's Tamil Censorship Highlights Risks to Everyone.

Bird, S. 2020. Decolonising speech and language technology. In *28th International Conference on Computational Linguistics, COLING 2020*, 3504–3519. online: Association for Computational Linguistics (ACL).

Bird, S. 2022. Local languages, third spaces, and other high-resource scenarios. In *60th Annual Meeting of the Association for Computational Linguistics, ACL 2022*, 7817–7829. Dublin: Association for Computational Linguistics (ACL).

Braun, V.; and Clarke, V. 2006. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2): 77.

Centre, B. . H. R. R. 2024. Dismantling the facade: A global south perspective on the state of engagement with tech companies.

Chang, T. A.; Arnett, C.; Tu, Z.; and Bergen, B. K. 2023. When is multilinguality a curse? language modeling for 250 high-and low-resource languages.

Christodouloupoulos, C.; and Steedman, M. 2015. A massively parallel corpus: the bible in 100 languages. *Language resources and evaluation*, 49: 375–395.

Coleman, D. 2018. Digital colonialism: The 21st century scramble for Africa through the extraction and control of user data and the limitations of data protection laws. *Michigan Journal of Race and Law*, 24: 417–439.

Commission, E. 2025. Commission adopts delegated act on data access under the Digital Services Act.

Couldry, N.; and Mejias, U. A. 2019. Data colonialism: Rethinking big data's relation to the contemporary subject. *Television & New Media*, 20(4): 336–349.

Das, D.; Guha, S.; Brubaker, J. R.; and Semaan, B. 2024. The "Colonial Impulse" of Natural Language Processing: An Audit of Bengali Sentiment Analysis Tools and Their Identity-based Biases. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. New York, USA: ACM.

Das, M.; Saha, P.; Mathew, B.; and Mukherjee, A. 2022. Hatecheckhin: Evaluating hindi hate speech detection models.

De Gregorio, G.; and Stremlau, N. 2023. Inequalities and content moderation. *Global Policy*, 14(5): 870–879.

Devi, V. S.; Kannimuthu, S.; and Madasamy, A. K. 2024. The Effect of Phrase Vector Embedding in Explainable Hierarchical Attention-Based Tamil Code-Mixed Hate Speech and Intent Detection. *IEEE Access*, 12(0): 11316–11329.

Divon, T.; and Ong, J. C. 2025. Tech Bro Power Play: Zuckerberg vs. Global Tech Justice.

Dourish, P.; and Mainwaring, S. D. 2012. Ubicomp's colonial impulse. In *Proceedings of the 2012 ACM conference on ubiquitous computing*, 133–142. New York, USA: ACM.

Elizabeth Dwoskin, J. W.; and Cabato, R. 2019. Content moderators at YouTube, Facebook and Twitter see the worst of the web — and suffer silently.

Elswah, M. 2024a. Moderating Kiswahili Content on Social Media.

Elswah, M. 2024b. Moderating Maghrebi Arabic Content on Social Media.

Errington, J. 2007. *Linguistics in a colonial world: A story of language, meaning, and power*. John Wiley & Sons.

Fanon, F. 2023. Black skin, white masks. In *Social theory re-wired*, 355–361. United Kingdom: Routledge.

Fine, M. 1994. Working the hyphens. *Handbook of qualitative research*, 2.

Fishman, J. A. 1989. *Language and ethnicity in minority sociolinguistic perspective*. United Kingdom: Multilingual Matters.

Garfinkel, S. L.; and Bowen, C. M. 2022. Preserving Privacy While Sharing Data.

Garimella, K.; and Chauchard, S. 2024. WhatsApp Explorer: A Data Donation Tool To Facilitate Research on WhatsApp.

Geiger, R. S.; Yu, K.; Yang, Y.; Dai, M.; Qiu, J.; Tang, R.; and Huang, J. 2020. Garbage in, garbage out? do machine learning application papers in social computing report where human-labeled training data comes from? In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, 325–336. New York, NY, USA: ACM.

Ghosh, S.; and Caliskan, A. 2023. ChatGPT Perpetuates Gender Bias in Machine Translation and Ignores Non-Gendered Pronouns: Findings across Bengali and Five other Low-Resource Languages. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, 901–912. New York, NY, USA: ACM.

Gorwa, R. 2019. What is platform governance? *Information, communication & society*, 22(6): 854–871.

Gramsci, A. 2020. Selections from the prison notebooks. In *The applied theatre reader*, 141–142. New York, USA: Routledge.

Gupta, S. 2024. The AI arms race: Which LLMs are winning the enterprise battlefield?

Haraway, D. 2013. Situated knowledges: The science question in feminism and the privilege of partial perspective 1. *Women, science, and technology*, 1: 455–472.

Held, W.; Harris, C.; Best, M.; and Yang, D. 2023. A material lens on coloniality in nlp.

Heller, M.; and McElhinny, B. 2017. *Language, capitalism, colonialism: Toward a critical history*. Canada: University of Toronto Press.

IDRC. 2024. Artificial Intelligence for Development.

Irani, L.; Vertesi, J.; Dourish, P.; Philip, K.; and Grinter, R. E. 2010. Postcolonial computing: a lens on design and development. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, 1311–1320. New York, NY, USA: ACM.

Iyer, P. 2025. What a New Study Reveals About Content Moderation in Tigray.

Jiménez, J. 2024. Worried About Meta Using Your Instagram to Train Its A.I.? Here's What to Know.

Kak, A. 2020. "The Global South is everywhere, but also always somewhere": National Policy Narratives and AI Justice. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, 307–312. New York, NY, USA: ACM.

Kapelke, C. 2020. Using differential privacy to harness big data and preserve privacy.

Kennedy, W. M.; and Campos, D. V. 2025. Vernacularizing Taxonomies of Harm is Essential for Operationalizing Holistic AI Safety. In *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '25, 698–710. New York, NY, USA: ACM.

Khan, M.; Ullah, K.; Alharbi, Y.; Alferaidi, A.; Alharbi, T. S.; Yadav, K.; Alsharabi, N.; and Ahmad, A. 2023. Understanding the research challenges in low-resource language and linking bilingual news articles in multilingual news archive. *Applied Sciences*, 13(15): 8566.

Kolli, V. 2024. Linguistic Colonialism: Moroccan Education and its Dark Past.

Kreutzer, J.; Caswell, I.; Wang, L.; Wahab, A.; van Esch, D.; Ulzii-Orshikh, N.; Tapo, A.; Subramani, N.; Sokolov, A.; Sikasote, C.; et al. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10: 50–72.

Kwet, M. 2019. Digital colonialism: US empire and the new imperialism in the Global South. *Race & Class*, 60(4): 3–26.

Legon, A.; and Alsalman, A. 2020. How Facebook can Flatten the Curve of the Coronavirus Infodemic. Technical report, Avaaz.

Lewis, J. E.; Abdilla, A.; Arista, N.; Baker, K.; Benesiinaabandan, S.; Brown, M.; Cheung, M.; Coleman, M.; Cordes, A.; Davison, J.; Duncan, K.; Garzon, S.; Harrell, D. F.; Jones, P.-L.; Kealiikanakaoleohaililani, K.; Kelleher, M.; Kite, S.; Lagon, O.; Leigh, J.; Levesque, M.; Mahelona, K.; Moses, C.; Nahuewai, I. I.; Noe, K.; Olson, D.; Parker Jones, Ō.; Running Wolf, C.; Running Wolf, M.; Silva, M.; Fragnito, S.; and Whaanga, H. 2020. Indigenous Protocol and Artificial Intelligence Position Paper. Project Report 10.11573/spectrum.library.concordia.ca.00986506, Aboriginal Territories in Cyberspace, Honolulu, HI.

Malik, S. 2022. Global labor chains of the western AI.

Mehta, I. 2023. X updates its terms to ban crawling and scraping.

MeitY. 2023. The Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021.

Milmo, D. 2021. Rohingya sue Facebook for £150bn over Myanmar genocide.

Mohamed, S.; Png, M.-T.; and Isaac, W. 2020. Decolonial AI: Decolonial theory as sociotechnical foresight in artificial intelligence. *Philosophy & Technology*, 33: 659–684.

Mufwene, S. S. 2004. *The ecology of language evolution*. United Kingdom: Cambridge University Press.

Nicholas, G.; and Bhatia, A. 2023. Toward Better Automated Content Moderation in Low-Resource Languages. *Journal of Online Trust and Safety*, 2(1).

Nicholas, G.; and Thakur, D. 2022. Learning to Share: Lessons on Data-Sharing from Beyond Social Media.

Nigatu, H. H.; and Raji, I. D. 2024. "I Searched for a Religious Song in Amharic and Got Sexual Content Instead": Investigating Online Harm in Low-Resourced Languages on YouTube. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 141–160. New York, NY, USA: ACM.

Nigatu, H. H.; Tonja, A. L.; Rosman, B.; Solorio, T.; and Choudhury, M. 2024. The Zeno's Paradox ofLow-Resource'Languages.

Obi-Young, O. 2018. Bantu's Swahili, or How to Steal a Language from Africa — Kamau Muiga.

Ògúnrẹ̀mí, T.; Nekoto, W. O.; and Samuel, S. 2023. Decolonizing nlp for "low-resource languages": Applying abebe birhane's relational ethics.

OpenAI. 2024. OpenAI and Reddit Partnership.

Ovalle, A.; Subramonian, A.; Gautam, V.; Gee, G.; and Chang, K.-W. 2023. Factoring the Matrix of Domination: A Critical Review and Reimagination of Intersectionality in AI Fairness. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, 496–511. New York, NY, USA: ACM.

Parrish, A.; Prabhakaran, V.; Aroyo, L.; Díaz, M.; Homan, C. M.; Serapio-García, G.; Taylor, A. S.; and Wang, D. 2024. Diversity-Aware Annotation for Conversational AI Safety. In Dinkar, T.; Attanasio, G.; Cercas Curry, A.; Konstas, I.; Hovy, D.; and Rieser, V., eds., *Proceedings of Safety4ConvAI: The Third Workshop on Safety for Conversational AI @ LREC-COLING 2024*, 8–15. Torino, Italia: ELRA and ICCL.

Perez, S. 2024. Reddit locks down its public data in new content policy, says use now requires a contract.

Perrigo, B. 2023. The Workers Behind AI Rarely See Its Rewards. This Indian Startup Wants to Fix That.

Popli, N. 2021. The 5 Most Important Revelations From the 'Facebook Papers'.

Posada, J. 2021. The Coloniality of Data Work in Latin America. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, 277–278. New York, NY, USA: ACM.

Quijano, A. 2000. Coloniality of power and Eurocentrism in Latin America. *International sociology*, 15(2): 215–232.

Quijano, A. 2007a. Coloniality and modernity/rationality. *Cultural studies*, 21(2-3): 168–178.

Quijano, A. 2007b. Questioning "race". *Socialism and democracy*, 21(1): 45–53.

Radiya-Dixit, E.; and Bogen, M. 2024. Beyond English-Centric AI Lessons on Community Participation from Non-English NLP Groups.

Rananga, S.; Isong, B.; Modupe, A.; and Marivate, V. 2024. Misinformation Detection: A Review for High and Low-Resource Languages. *Journal of Information Systems and Informatics*, 6(4): 2892–2922.

Rowe, J. 2022. Marginalised languages and the content moderation challenge.

Said, E. W. 1977. Orientalism. *The Georgia Review*, 31(1): 162–206.

Said, E. W. 2000. *Out of Place—A Memoir*. United Kingdom: Vintage Books.

Sambasivan, N.; Kapania, S.; Highfill, H.; Akrong, D.; Paritosh, P.; and Aroyo, L. M. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. New York, USA: ACM.

Samuels, E. 2020. How misinformation on WhatsApp led to a mob killing in India.

Sap, M.; Card, D.; Gabriel, S.; Choi, Y.; and Smith, N. A. 2019. The Risk of Racial Bias in Hate Speech Detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1668–1678. Florence, Italy: Association for Computational Linguistics.

Scarcella, M. 2024. Elon Musk's X wins appeal to block part of California content moderation law.

Scheuerman, M. K.; and Brubaker, J. R. 2024. Products of Positionality: How Tech Workers Shape Identity Concepts in Computer Vision. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24. New York, NY, USA: ACM.

Schöpf, C. M. 2020. The Coloniality of Global Knowledge Production: Theorizing the Mechanisms of Academic Dependency. *Social Transformations: Journal of the Global South*, 8(2): 5–46.

Schwartz, L. 2022. Primum Non Nocere: Before working with Indigenous data, the ACL must confront ongoing colonialism. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 724–731. Dublin, Ireland: Association for Computational Linguistics.

Shahid, F.; and Vashistha, A. 2023. Decolonizing Content Moderation: Does Uniform Global Community Standard Resemble Utopian Equality or Western Power Hegemony? In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. New York, USA: ACM.

Siapera, E. 2022. AI Content Moderation, Racism and (de) Coloniality. *International Journal of Bullying Prevention*, 4(1): 55–65.

Sreelekha, S.; Bhattacharyya, P.; and Malathi, D. 2018. Statistical vs. rule-based machine translation: A comparative study on indian languages. In *International Conference on Intelligent Computing and Applications: ICICA 2016*, 663–675. Australia: Springer.

Stokel-Walker, C. 2024. Under Elon Musk, X is denying API access to academics who study misinformation.

stream. 2025a. Audio Moderation.

stream. 2025b. Video Moderation.

Thakur, D. 2025. Moderating Quechua Content on Social Media. Technical report, Center for Democracy and Technology.

Thiong'o, N. u. i. w. 1986. *Decolonising the Mind: The Politics of Language in African Literature*. East Africa: EAEP.

TikTok. 2024. Supporting independent research.

Udupa, S.; Maronikolakis, A.; and Wisiorek, A. 2023. Ethical scaling for content moderation: Extreme speech and the (in) significance of artificial intelligence. *Big Data & Society*, 10(1): 1–15.

Union, U. G. 2025. Content moderators launch first-ever global alliance, demand safe working conditions and accountability from tech giants.

van Esch, D.; Sarbar, E.; Lucassen, T.; O'Brien, J.; Breiner, T.; Prasad, M.; Crew, E.; Nguyen, C.; and Beaufays, F. 2019. Writing across the world's languages: Deep internationalization for Gboard, the Google keyboard.

Varshney, K. R. 2024. Decolonial AI Alignment: Openness, Viśesa-Dharma, and Including Excluded Knowledges. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7 of *AIES '24*, 1467–1481. New York, NY, USA: ACM.

Verran, H.; and Christie, M. 2007. Using/designing digital technologies of representation in Aboriginal Australian knowledge practices. *Human Technology*, 3(2): 214–227.

Villenas, S. 1996. The colonizer/colonized Chicana ethnographer: Identity, marginalization, and co-optation in the field. *Harvard educational review*, 66(4): 711–732.

Wei, J. T.-Z.; Zufall, F.; and Jia, R. 2025. Operationalizing Content Moderation "Accuracy" in the Digital Services Act. In *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '25, 1527–1538. New York, NY, USA: ACM.

Witness, G. 2022. Facebook unable to detect hate speech weeks away from tight Kenyan election.

Wong, J. C.; and Ernst, J. 2021. Facebook knew of Honduran president's manipulation campaign – and let it continue for 11 months.

Wong, J. C.; and Harding, L. 2021. 'Facebook isn't interested in countries like ours': Azerbaijan troll network returns months after ban.

Yibeltal, K.; and Muia, W. 2023. Facebook's algorithms 'supercharged' hate speech in Ethiopia's Tigray conflict.

Zevallos, R.; and Bel, N. 2023. Hints on the data for language modeling of synthetic languages with transformers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 12508–12522. Toronto, Canada: Association for Computational Linguistics.

Zhong, T.; Yang, Z.; Liu, Z.; Zhang, R.; Liu, Y.; Sun, H.; Pan, Y.; Li, Y.; Zhou, Y.; Jiang, H.; et al. 2024. Opportunities and challenges of large language models for low-resource languages in humanities research.